Ecological Adaptation in the Context of an Actor-Critic

Ignasi Cos-Aguilera

Doctor of Philosophy
Institute of Perception, Action and Behaviour
School of Informatics
University of Edinburgh
2005

Abstract

Biological beings are the result of an evolutionary and developmental process of adaptation to the environment they perceive and where they act. Animals and plants have successfully adapted to a large variety of environments, which supports the ideal of inspiring artificial agents after biology and ethology. This idea has been already suggested by previous studies and is extended throughout this thesis. However, the role of perception in the process of adaptation and its integration in an agent capable of acting for survival is not clear.

Robotic architectures in AI proposed throughout the last decade have broadly addressed the problems of *behaviour selection*, namely deciding "what to do next", and of *learning* as the two main adaptive processes. Behaviour selection has been commonly related to theories of motivation, and learning has been bound to theories of reinforcement. However, the formulation of a general theory including both processes as particular cases of the same phenomenon is still an incomplete task. This thesis focuses again on behaviour selection and learning; however it proposes to integrate both processes by stressing the ecological relationship between the agent and its environment. If the selection of behaviour is an expression of the agent's motivations, the feedback of the environment due to behaviour execution can be viewed as part of the same process, since it also influences the agent's internal motivations and the learning processes via reinforcement. I relate this to an argument supporting the existence of a common neural substrate to compute motivation and reward, and therefore relating the elicitation of a behaviour to the perception of reward resulting from its execution.

As in previous studies, behaviour selection is viewed as a competition among parallel pathways to gain control over the agent's actuators. Unlike for the previous cases, the computation of every motivation in this thesis is not anymore the result of an additive or multiplicative formula combining inner and outer stimuli. Instead, the ecological principle is proposed to constrain the combination of stimuli in a novel fashion that leads to adaptive behavioural patterns. This method aims at overcoming the intrinsic limitations of any formula, the use of which results in behavioural responses restricted to a set of specific patterns, and therefore to the set of ethological cases they can justify. External stimuli and internal physiology in the model introduced in this thesis are not combined a priori. Instead, these are viewed from the perspective of the agent as modulatory elements biasing the selection of one behaviour over another guided by the reward provided by the environment, being the selection performed by an actor-critic reinforcement learning algorithm aiming at the maximum cumulative reward.

In this context, the agent's drives are the expression of the deficit or excess of internal resources and the reference of the agent to define its relationship with the environment. The schema to learn object affordances is integrated in an actor-critic reinforcement learning algorithm, which is the core of a motivation and reinforcement framework driving behaviour selection and learning. Its working principle is based on the capacity of perceiving changes

in the environment via internal hormonal responses and of modifying the agent's behavioural patterns accordingly. To this end, the concept of reward is defined in the framework of the agent's internal physiology and is related to the condition of physiological stability introduced by Ashby, and supported by Dawkins and Meyer as a requirement for survival. In this light, the definition of the reward used for learning is defined in the physiological state, where the effect of interacting with the environment can be quantified in an ethologically consistent manner.

The above ideas on motivation, behaviour selection, learning and perception have been made explicit in an architecture integrated in an simulated robotic platform. To demonstrate the reach of their validity, extensive simulation has been performed to address the affordance learning paradigm and the adaptation offered by the framework of the actor-critic. To this end, three different metrics have been proposed to measure the effect of external and internal perception on the learning and behaviour selection processes: the performance in terms of flexibility of adaptation, the physiological stability and the cycles of behaviour execution at every situation. In addition to this, the thesis has begun to frame the integration of behaviours of an appetitive and consummatory nature in a single schema. Finally, it also contributes to the arguments disambiguating the role of dopamine as a neurotransmitter in the Basal Ganglia.

Acknowledgements

I would like to first thanks my first supervisor Gillian Hayes. Gill has guided me along these years with valuable suggestions and criticisms, always presented in a scientific and delicate manner. Thank you also for helping me put the pieces of the puzzle together in a coherent and sensible story and for devoting as much time correcting its expression. Also many thanks for being in the dark hours.

Thanks also to my second supervisor Lola Cañamero, for finding time and resources to continue this PhD and for offering me the opportunity of enjoying an academic and personal experience at the University of Hertfordshire. Thank you also for providing a different inspiration and perspective to the formulation of the thesis and for encouraging me to publish some parts of it. I am also deeply thankful for helping me out with the correction of the thesis.

I want to specially thank Andrew Gillies, my other second supervisor, for being a constant guidance and an excellent companion. Thank you for imposing logic and common sense when other emotions were not good leaders and for providing the best guidance from a neuroscience viewpoint. Furthermore, thank you for being always available and for sharing interesting discussions with me. Also thanks for suggesting appropriate changes to my pattern of behaviour when stress has been beyond every threshold. Furthermore, thank you for bringing me closer to a mix of Scottish and Maori culture and sense of humor, and for sharing with me the taste of cultural affinity and of friendship. For all of this I am deeply grateful.

I also want to thank Alan Smaill, for helping me with the initial formulation of this thesis, and to Bob Fisher for his good advice and for generously funding several trips to some conferences (through the IPAB funds for students). This thesis would not have been possible without this.

The time of this PhD has been a time of substantial changes in my life and in the academic environment where I developed this thesis. However, there has been always a friendly and warm atmosphere at Forrest Hill and at King's Buildings, with the IPAB and the IANC, as well as with the STRC at the University of Hertfordshire. Thanks to all of you.

During these years there has been some special friends who have encouraged me to continue this PhD and who kept reminding me that submission was every time closer! The same who have taken a constant interest in me at an academic and at a personal level. I want to specially mention José Carmena and George Konidaris for being there, even when stress is the only constant. Furthermore, I also want to thank George Maistros, Yuval Marom and Aroosha Laghee for making of the IPAB a friendly and warm space for scientific discussion. Thanks also to Jay Bradley, Paul Crook, Tim Lurkins and Pete Ottery for the valuable discussions and for their companionship. I also want to thank Moray Alan, Joanna Young, John Quinn, Wolfgang Lehrach and Lena Hannson for turning my intrusion in the IANC into a friendly and enjoyable time. Also special thanks to David Willshaw and Chris Williams for making this

intrusion possible and to Emma Black for being always herself. I also want to specially mention Orlando Ávila-García for his friendly hospitality, his valuable advice and good judgment regarding robot emotions and Ben Robins for his lessons of life. Thanks very much to both of you.

Also thanks very much to Lizelle Bisschoff-Minnaar for sharing her artistic sense with me, for her availability and for sharing with me the very many ways of to enjoy life and to Marina Papoutsi for demonstrating that constant joy of life does not belong to a world of phantasy. Thanks very much for your love and support. Very special thanks for Louis Atallah, for being a very special companion and friend at any reasonable and unreasonable time and for helping me out with in any way he could with this thesis. Very many thanks to Miguel A. Cazorla at the University of Alicante for being an informal advisor, for sharing personal interests and for always keeping a space for discussion and advice within an impossible agenda. Thanks to you and to Amparo for all your friendship and support.

I also want to specially thank Theodoros (Akis) Damoulas for having faith in some of my ideas and for turning them into a brilliant MSc thesis. Also many thanks for adding good doses of mediterranean humor to our working environment.

There are also many people outside the working environment who have also provided valuable help and motivation. I feel very lucky for having been able to share these years with Manolis Sifalakis and with Yves-Marie Legrand (you are really the great one!) for keeping always a constant eye on me when things were right or wrong and for providing me with many hours of fun, friendship and always helpful advice.

Thanks very much to Benedetta and to Dimitrios for their welcome to London and for their always warm hospitality and good friendship. Very many thanks to Arturo Sanz and Cristina Piqueras for their hospitality, support and love during an initially brief, but colourful time of our lives.

In Barcelona I also have to thank Silvia Turchin for her friendship and companionship during this time, and Katleen Koncz for her love and support during this time. I am very much indebted with Katerina Papachroni, for her friensdhip, sweetness, patience and understanding in sometime very complicated situations. Also many thanks to Fantina Madricardo for her friendship and for believing and supporting me in the dark hours. Also many thanks to Pier Paracchini and Silvia Spoletini for keeping a constant interest in my life.

I cannot end this list of thanks without mentioning two of the most important persons in my life: Alexo Esperato and Eduard Lax. I am grateful beyond imagination for your support, patience, advice, friendship, crazyness and humor at any time of day or night during these past years and I hope many more to come.

Lastly, this would not have been possible without the love and support from my family; despite losses and disagreements. I am deeply indebted to my mother Mariana, who did not

stop loving and believing in me to the end, and for having left such a sensible and beautiful legacy. Massive thanks to my father Pau, for being the very constant of my life and for being an example of good will, honesty, persistence, availability and love. Thanks to my brother and my sister, Josep Manel and Anna, just for everything. Also thanks to my aunt Mary J. and my cousin Amèlia and to the rest of my family who supported me beyond any reasonable understanding.

To my mother Mariana, deeply loved, dearly missed.

Table of Contents

List of Figures

List of Tables

Chapter 1

Learning Affordances and Behavioural Patterns with an Actor-Critic

Artificial Intelligence does not anymore view robots as pieces of hardware executing a plan, but as creatures existing in an environment they perceive and wherein they act. Instead of plans and reasoning mechanisms, robots decide their course of action based upon their internal representations and perception of their environment. The advantage of this second approach is that the robot does not need to be programmed; the robot can by itself design its own plan to perform a task. However, the fact that these robots do not require a completely specified program does not mean they are easy to design. If you ever tried to program a robot to perform a task, you already know the complexity of this endeavour. We are all capable of predicting the effect of some simple interactions with the environment, but predicting the consequences of every single action is not possible. Hence, lots of fine tuning may be necessary before the robot performs as intended. In response to this, several authors have addressed different aspects of this problem. These have provided very useful insights into the role of affective phenomena for behaviour selection (Ávila-García and Cañamero, 2004), an ethological model of motivation based on internal physiology (Spier and McFarland, 1996), a description of the procedures used by some animals to select behaviours (Gurney et al., 1998) and some opinions on the elements that would suffice to build animated creatures (Blumberg, 1997). However, these have also highlighted two main deficiencies. Firstly, that perception is disregarded as an adaptive mechanism; secondly, that for most ethological models, once behavioural responses are set they are difficult to change. My intention in this context is to review perception as an adaptive mechanism and to integrate this into an adaptive agent architecture. To attain this goal, I have searched for inspiration in previous studies on adaptation based on ethology, neuroscience and robotics.

1.1 The Problem

Therefore, the problem is to find the appropriate manner to introduce perception into the dynamics of interaction with the environment. To this end I have searched for inspiration in neuroscience and ecology. The last fifteen years have consolidated biology as a source of inspiration to build robotic models. An example of this is the model of action selection proposed by Redgrave, Prescott and Gurney (Gurney et al., 1998), based on the assumption that the main role of a vertebrate's basal ganglia is to arbitrate among the animal's behaviours in a centralised manner (Redgrave et al., 1999).

From a neuroscience perspective, I agree that the basal ganglia play a significant role in action selection. However, this model does not explain the process of adaptation to the environment, since the meaning of every object is engineered and hard-coded before the selection of behaviours initiates. Therefore, the execution of behaviours will ultimately be unique and completely determined by these definitions. This may suffice in a static environment for some applications. However, interacting in a dynamic environment will —where the value of an object may vary throughout time— require learning of these values. As a response to this, I argue that it is necessary to modify the designer's view with respect to the role of the environment. The environment has already been demonstrated to provide proprioceptive and kinaesthetic feedback when executing a behaviour. However, solving the problem of perceiving in a dynamic environment necessitates a change of perspective with regard to the environment. The value of an object is used to bias the execution of one behaviour over another. Hence, there is an implicit behavioural meaning, which I suggest to relate to the notion of affordance. According to this, the environment should not be viewed anymore as a set of sensory signals, but as a set of potentialities for action, which may vary over time. This novel view of the environment introduces a direct relationship between objects in the environment and their potentialities for action, which is equivalent to a basic semantic definition of the surrounding environment with regard to the agent. However, it also introduces the problem of learning them. Redgrave's model uses situatedness for action selection. I argue that this can be further extended in the concepts of motivation and internal physiology developed by Spier, McFarland and Canamero (Spier and McFarland, 1996; Cañamero, 1997) and the role of reward introduced by Rolls (2003).

Therefore, the learning of affordances can be formalised by relating the effect of executing a behaviour to the agent's internal physiology. For example, if an object is edible, eating it should have a compensatory effect on the level of hunger of the agent. If this effect occurs repeatedly, it can be assumed that the object is always edible. Furthermore, if the effect of that particular object varies, the internal representation of this semantic will also vary, suggesting appropriate changes in the agent's behavioural patterns. Along the same lines, Spier and McFarland's motivational model (Spier and McFarland, 1996), introduced the possibility

of combining stimuli with internal physiology to suggest the selection of one behaviour over another. The intensity of each behaviour is calculated on the basis of its related motivation, which multiplies the intensity of a stimulus by its related internal physiological values. This model is mostly inspired by ethology and is a referent with regard to later models of motivation. However, this model fails at explaining behavioural phenomena, since the combination of stimuli is restricted a priori. I suggest that reward, as an assessment of an interaction with the environment, can be the solution to this problem. I adhere to Schultz' hypothesis that the basal ganglia are a learning device (Schultz et al., 1993). This argues that the basal ganglia are a device learning to relate stimuli to responses on the basis of an assessment signal (reward), which I argue is also reinforcing or weakening the tendencies towards one behavioural response or another. This is therefore suggestive of a procedure for learning to select behaviours, which I propose to implement in a model.

1.2 The Thesis

This thesis focuses on two *mechanisms* for an agent to dynamically *adapt* to an environment: *learning* the *affordances*[1] of the elements of the environment surrounding the agent and *learning to sequence the executions of behaviours* to survive. Although these topics have been already addressed, this thesis proposes a novel approach inspired after the ecological principle. The agent and the environment are integrated in a single dynamics; therefore, changes of the environment reflect in changes of in the agent. Based on this, any living being and its environment are viewed as parts of a single entity, governed by a common dynamics. Therefore, from an artificial perspective, changes of the environment should reflect in changes in the relationship between the agent and the environment. I have considered *perception*, *behaviour selection* and *learning* as three of the necessary processes coupling the agent and its environment.

Firstly, as sources of inspiration for *perception*, this work has focused on the notion of *affordance* formulated by Gibson (1966). In a way, perception in robotics has been often treated as a set of sequential processes, e.g., the recognition of individual features often precedes the recognition of other more elaborated features (composed from the individual features), which are processed to govern the agent's behaviours. In contrast, the implementation introduced in this thesis views perception from a functional viewpoint. The agent perceives a flow of sensory patterns while interacting with its environment, and *learns* to assign a predictive value to some patterns depending on previous experience. If the perception of a certain situation preceded the successful execution of a certain behaviour, the future perception of a similar pattern will be used as a predictor of the potentiality of performing that same behaviour. This relational concept between the perception of certain elements of the environment and the potentiality of

[1] Affordance is a relational concept meaning the function offered by the objects in the environment surrounding the agent. This is further explained in chapter 4.

performing an action is the affordance. Learning affordances is viewed as the process relating perception to the agent's physiology. From a different perspective, this also means that this learning process defines the semantics of the objects in the environment with regard to that particular agent as a natural result of their intrinsic common dynamics. In other words, learning affordances implies grounding the agent in the environment (Harnad, 1990).

Nevertheless, adaptation often requires more than learning the potentialities for action offered by an environment. Knowing the affordances of the objects in the surrounding environment only allows the agent to react to them. However, the next action to perform is underdetermined when the goal is survival in a competitive or dynamic environment or when the object nearby offers more than one course of action. In order to address this problem, I have taken advantage of further biological inspiration by integrating the affordance learning system into an adaptive system based on the actor-critic algorithm (Sutton and Barto, 1981). This element of design learns behavioural patterns. It has been selected because it conciliates the traditional ethological view on action selection brought up by Avila-García and Cañamero (2002); Cañamero (1997); Blumberg (1997); Tyrrell (1993) with the neuroscience and machine learning perspective, which views learning and behaviour selection as concurrent processes (Konidaris, 2003; Humphries, 2002; Dayan and Balleine, 2002; Humphrys, 1997). I argue that both views can be conciliated if the choice of behaviours is driven by the maximisation of future reward. This has been addressed by the model described in chapter 5, aiming at demonstrating that an agent driven by reward can learn behavioural patterns coherent with ethological observation. Furthermore, by assuming this principle, it is also possible to view behaviour selection and learning as concurrent processes in the hierarchy of adaptation (see the beginning of 2). Behaviour selection provides a certain level of adaptiveness, which can only be improved if processes ranking higher in the hierarchy, such as learning, come into play. However, this view is incomplete if these processes are considered to be independent. Both these processes, behaviour selection and learning, are related to the same dynamics of which perception is also part. Hence, I argue that both must be considered concurrently. The agent's internal motivations are the expression of internal resources and of its perception of the environment, hence motivations are the reference for choosing a behaviour to relate to the environment. These considerations are the framework of this thesis, which focuses on the study of learning affordances and behavioural patterns in a situated ecological agent.

Contributions

The use of an ecological approach has provided some answers to problems posed in the traditional behaviour selection architectures. Former studies in robotics are mostly based on selection architectures where the intensity of the motivations or drives biasing the execution of one behaviour or another is calculated via additive and/or multiplicative formulae. Examples of

this are Avila-García and Cañamero (2002); McFarland and Spier (1997); Spier and McFarland (1996). These exhibit two main limitations; they do not explain some behavioural situations (McFarland, 1993), furthermore they do not generally consider new behavioural patterns can be modified to deal with changing situations. The model proposed in this thesis has overcome these limitations by using the ecological principle to constrain this relationship. Therefore, variations in the availability and distribution of resources in the environment will be perceived by the agent, which will respond by appropriately modifying its behavioural patterns. I have formulated our behaviour selection architecture by extending the *hypothesis* of an actor-critic reinforcement learning algorithm to drive the learning of stimulus-response relationships in Pavlovian and in instrumental contingencies. This hypothesis is based on neurological evidence for Pavlovian contingencies only (Schultz et al., 1993), although it has also been suggested for the instrumental case (Houk et al., 1995). This learning architecture assumes that learning and behaviour can be integrated if decisions are made by comparing the predictions of future return for the execution of every behaviour. If the prediction does not match the real reward obtained after the execution of the behaviour, this is corrected for future interactions. The policies for selecting behaviours are appropriately modified, and the potential of executing that behaviour when those patterns are perceived are also adapted. The results of testing the extended version of the actor-critic hypothesis have a dual effect. Firstly, they integrate behaviour selection and learning in a single ecological framework; secondly, this helps to disambiguate the role of dopamine (DA) as a neuro-transmitter in the basal ganglia (Redgrave et al., 1999) supporting its role as an assessment signal for Pavlovian and instrumental learning. This is further explained in the next chapter.

This thesis is not explicitly addressing planning (Chapman, 1987) or the modulation of behavioural responses due to affective phenomena (Fellous, 2004). However, I understand that these also play a role in adaptation. This thesis mainly focuses on the role and reach that ecological learning has in the context of an agent embodying affordances and an actor-critic algorithm to adapt to its environment.

In summary, the main contributions of this thesis are:

- An *ecological adaptive agent* that integrates perception and behaviour selection in a single framework. Perception is structured in the form of affordances and behaviour selection on the structure and principles of the actor-critic. The perception and the behaviour selection architectures improve the agent adaptation by interacting with the environment (via correlating the fluctuations of the agent's internal physiology to the perception of the agent's itself, i.e., according to the ecological principle). These architectures are based on integrating interaction and internal physiology in the same dynamics.

- The ecological principle constrains the manner in which to interact with the environment.

This principle I have also used to inspire the way in which *stimuli are combined.* Based on the actor-critic, the agent overcomes the limitations when stimuli are combined using additive or multiplicative formulae. Instead, it provides a more complex account of the influence of external affordances and internal drives and of the ways in which these combine to interact with one another to provide behavioural patterns leading to an internal physiological stability.

- In terms of *perception*, this thesis has provided a concrete methodology to cluster sensory signals and to build a neural representation the object's affordances with respect to the agent's internal physiology. The learning is based on the causality between physiological fluctuations and the execution of behaviours when certain sensory patterns are perceived. It has been experimentally demonstrated that these principles and the model introduced suffice to build a perception space for the agent to successfully interact with a variety of scenarios in which it adapts in a dynamic manner.

- With regard to *behaviour selection*, an architecture based on an actor-critic as a phenomenological model of the basal ganglia has been introduced. This architecture has experimentally demonstrated its ability to combine external and internal stimuli to give rise to behavioural patterns which satisfy the need for internal physiological stability postulated for the agent's survival.

  - It proposes to measure the effect of behaviour execution in terms of reward and relates this to homeostasis and to the internal physiological stability. Related to this, the *relationship between the effect due to the execution of a behaviour and its related reward* has been hypothesised. Reward is the metric used to assess the performance of the actor-critic and therefore the appropriateness of the behavioural patterns for the given scenario. A formula that relates the effect of executing a behaviour to a certain amount of reward has been proposed based on ethological data. If the effect of the behaviour diminishes an internal deficit, a reward results from the formula. Conversely, a punishment is obtained. This implicitly relates the maximisation of reward to obtaining internal physiological stability.

  - It introduces experimental evidence supporting the idea that the "common currency" used for comparing among several behavioural possibilities is the expectation of future reward.

- From a neurological perspective, it also helps to disambiguate the role of dopamine (DA) in the basal ganglia by supporting its hypothesis as the signal of error in the prediction of reward. Furthermore, it suggests that Pavlovian and instrumental learning could use the same assessment criterion and therefore share part of the same neural substrate.

- The thesis also demonstrates that the internal physiological dynamics plays a fundamental role in *grounding knowledge* from the environment and in relating the different processes encompassed in the adaptation process. Furthermore, the last experiments also demonstrate that the parameters of the *internal physiology modulate the learning* in order to divert most attention to those elements of the environment that palliate the deficits which grow faster. Therefore, the dynamics of the environment and of the internal physiology are inter-related.

- Finally, this whole set of principles has been explicitly integrated in a model that has demonstrated the ability to adapt the agent to the environment in an ecological manner. Furthermore, this has demonstrated that the integration between perception, behaviour selection and the environment can be merged in a common dynamics governed by the principle of ecological adaptation. This argument extends further in the discussion of this thesis, suggesting future ways of application in robotics.

1.3 The Organisation of the Thesis

The thesis addresses the problem of adaptation from an ecological perspective, and is divided into the following chapter:

Chapter 2 introduces a literature review, highlighting the need to encompass the different elements of the internal dynamics of the agent with the level of availability of resources and complexity of the scenario, and the advantage of doing so in an ecological manner.

Chapter 3 introduces an overview of the model and of the points to be addressed in an artificial model.

Chapter 4 introduces the homeostatic architecture and the module for learning object affordances from the objects in the scenario.

Chapter 5 presents the architecture for learning behavioural patterns, and the different elements of reinforcement learning used in order to learn the appropriate policies to reach physiological stability.

Chapter 6 studies the effect of internal modulation on the learning of behavioural patterns for a set of behaviours.

Chapter 7 is the Discussion.

Chapter 8 is the Conclusion.

Chapter 2

Literature Review

This thesis addresses the process of learning affordances and its integration in the adaptive process. This is inspired and framed by the ecological principle. In this light, I have studied *perception*, *behaviour selection* and *learning* as three of the necessary processes coupling the agent and its environment in a dynamic manner.

The environment is viewed as a set of potentialities of action (affordances), which will have to be intelligently selected and executed to survive. In order to frame this, the chapter introduces the ecological framework where the agent will have to make decisions (by appropriately combining external and internal stimuli). This precedes a description of the behaviour selection strategies required to learn to maintain the agent 'alive'.

I have assumed that the agent is situated in its environment and that as for biological beings this is necessary for survival. Therefore, the perception of potential courses of action and the appropriate sequencing of behaviours will have to lead towards the compensation of internal needs. In this process, the perception of affordances is viewed as the process of acquisition of basic knowledge from the environment, which situates or grounds the agent to its environment. In a complementary fashion, the actor-critic provides the necessary ability of sequencing the execution of behaviours to maximise reward. Depending on how reward is defined, the actor-critic biases the agent towards reactive or internally motivated behavioural patterns.

Therefore, the behavioural patterns that ensure the agent's survival will have to draw on Ashby's criterion of physiological stability (Ashby, 1965). Furthermore, we suggest basing our model on the *actor-critic* reinforcement learning algorithm, which has been hypothesised by Houk et al. (1995) to govern high level behaviour selection and instrumental learning in high vertebrates, such as rats and macaques.

Adaptation in the natural world occurs both at a genetic and developmental level. However, this thesis solely focuses on developmental learning in terms of reward as an adaptation mechanism. This chapter introduces background information regarding this. Furthermore, it has been considered appropriate to also review in this chapter social learning and learning by

imitation for reference purposes and for completing our view on biologically inspired learning processes (Demiris and Hayes, 2002; Maistros and Hayes, 2001; Billard and Mataric, 2001; Mataric, 2000).

This chapter introduces a view of perception, behaviour selection and learning as a continuum of adaptive processes in the context of ecological psychology, from social facilitation to taxis reactions. It highlights that the notion of interaction with the environment and the manner in which this happens is fundamental to condition behaviour selection and the task to be performed. Finally, it argues in favour of a framework based on the actor-critic reinforcement learning algorithm as a device to improve the agent's adaptation to the environment. The chapter is organised as follows:

- Section 2.1 introduces a *taxonomy of adaptive processes*, classifying them in terms of their time-scale.

- Section 2.2 introduces two fundamental processes for adaptation: *learning* and *behaviour selection*. These are characterised in terms of persistence (see section 2.2.2) and the manner in which stimuli are combined (see section 2.2.1).

- Section 2.3 introduces the necessary elements to abstract an *artificial physiology* as a framework in which to embed any adaptive system that has to provide adaption. Furthermore, it also presents the concept of motivation, which biases the agent to make decisions. This is used throughout the experimental chapters.

- This is followed by an introduction to *reinforcement learning* and of its relationship to behaviour selection and the definition of reward (see section 2.4).

- Section 2.5 introduces the *principles of neuroscience* and ethology that have inspired our model and that will condition the fashion in which the actor-critic combines external and internal stimuli and therefore the reach of its adaptiveness.

- Section 2.6 explains Gibson's *ecological perception* (Gibson, 1966) and the reasons to use this as a principle for artificial adaptive agents. This is further extended in sections 2.7 and 2.8, which introduce the neurological background involved in *affordance learning* and some theoretical considerations on *grounding* knowledge from the environment, respectively.

- Section 2.9 introduces the *complete framework* that will lead to building the model introduced in the next chapter and a summary presents the set of hypotheses to be tested in the experimental chapters.

2.1 Taxonomy of Adaptive Processes

Adaptation is about modifying behavioural patterns in response to changes in the environment to survive. Therefore, it seems reasonable to classify adaptive processes according to the time-scale of the changes they encompass (Maes, 1997). Namely:

- Behaviour Selection. This consists of a change of activity typically due to a change in the environment (e.g., a predator is perceived in an agent's vicinity) or to a fluctuation of the agent's internal physiology. The response to each contingency is usually unique. Some of these responses are the result of genetically encoded information and of developmental learning. Nevertheless, behaviour selection specifically refers to a change in activity from a behaviour to another, where each individual behavioural pattern remains fixed.

- Developmental Learning. This refers to changes within the patterns of behaviour and/or perception.[1] These changes occur throughout the life of the agent via interaction with its environment. This sort of developmental learning has been sub-divided into three different types:

  - Associative Learning. This consists of establishing relationships between stimuli and their elicited responses. An example of this is Stimulus-Response learning, formally introduced by Pavlov (1927).

  - Reinforcement Learning. This consists of relating stimuli to actions leading to reward. The learning of relationships requires more than one interaction of the same sort and can be unlearnt along the same lines. From a behavioural viewpoint, the term had already been used by Pavlov, though probably the most popular model is due to Sutton and Barto (1981). Thus since this learning method requires several interactions, it exhibits a longer time-scale than other learning schemata.

  - Social Learning. This comprises every adaptive sub-process, ranging from behaviour selection to the modification of behavioural patterns via interaction with other individuals. This includes social facilitation (Heyes and Galef, 1996), learning by imitation (Demiris and Hayes, 2002) and also overlaps with reinforcement learning for the cases in which the assessment (reward) is indirectly provided by demonstrators.

- Genetic Adaptation. This comprises changes in structure and behaviour of the individual transferred from generation to generation. These changes extend over several generations to have effect (Damoulas, 2004; Wright, 1932, 1931; Fisher, 1930; Darwin, 1866).

[1] If a particular pattern of perception means the potentiality of performing one behaviour or another.

The categories of adaptive processes described above take place concurrently in nature. However, in a strict sense only behaviour selection and developmental learning facilitate an animal's adaptation in a dynamic and interactive fashion within its life time. Similarly, we would like to model this process in an efficient manner to be applied to robotic platforms. To this end, this thesis focuses on developmental, associative and reinforcement learning as adaptive processes; deliberately disregarding genetic adaptation. Despite this process being intrinsically related to the ecological relationship between the agent and its environment, our interest only focuses on the adaptive processes at a lifetime scale for a single individual, and social learning involving several individuals is disregarded.

2.2 Behaviour Selection

Behaviour selection is the change of current activity in response to internal or external stimuli. Several authors have addressed this issue from different disciplines: ethology (Blumberg, 1997; Spier and McFarland, 1996; Rosenblatt and Payton, 1989; Baerends, 1976; Dawkins, 1976; Tinbergen, 1953), artificial intelligence (Gershenson García, 2000; González et al., 2000; Maes, 1991), psycho-neuroscience (Bryson, 2004; McClure et al., 2003; Bryson, 2000; Redgrave et al., 1999), machine learning (Sutton and Barto, 1998; Humphrys, 1997; Sutton and Barto, 1981) and robotics (Avila-García and Cañamero, 2002; Tyrrell, 1993).

The different overviews highlight the fact that selecting suitable behaviours is a complex task involving several related processes. Even for the simple case of a reactive behaviour, making a decision will require sufficient information from the environment. Furthermore if the decision is deliberative, appropriate methods for the combination of stimuli must be provided in order to satisfy some criteria of goodness with regard to the goals of the agent. Moreover, there are related aspects of the creature and its environment that control the execution of the behaviour, either before (anticipatory), during or after it. For example, how long should the execution of a behaviour be maintained? When should the agent change its activity? Do behavioural patterns have to exhibit intentionality and to be motivation driven? (Blumberg, 1994).

Among others, these issues have been formally listed in a complementary manner by Blumberg (1997), Maes (1997) and Tyrrell (1993). They are introduced in the next sub-section, except those perception related, introduced in section 2.6. Finally, section 2.9 addresses the missing links and the methodology followed throughout the thesis.

2.2.1 Combination of Stimuli

How can an agent combine its internal wishes with the possibilities for action provided by its close environment? This issue has been commonly referred to as *combining external and*

internal stimuli to satisfy internal goals (McFarland, 1993). Even though the problem has been widely addressed in several fields, models based on ethological observation focus most of our attention. These provide an analysis that correlates the agent's internal physiological dynamics with behavioural observations, which makes them most valuable to us. The validity of an artificial model can be assessed in an analogous manner and furthermore be compared with an ethological counterpart.

Building an artificial model requires the formulation of abstract representations mirroring the animals' internal physiology (Cañamero, 1997; Spier and McFarland, 1996). Related to this, the notion of *motivation* is fundamental (Damasio, 2000; Izard, 1993; Toates and Jensen, 1990). This is defined as a combination of internal and external stimuli expressing an urge for compensation. Internal stimuli were firstly abstracted by Hull in his concept of drive (Hull, 1943), expressing the urge to act whenever the level of an internal resource falls below a certain threshold, e.g., when feeling hungry, sated, weary, invigorated, thirsty. This information is then processed to select one behaviour over another. However, the decision making is a complex task, since there are as many right decisions as criteria defining correctness. To constrain this, ethological coherence has been traditionally applied, namely comparison with animal behaviour. As a result, several classical **procedures for the combining of stimuli** have been formulated. These are listed next.

- Hull (1943) formally introduced the concept of *drive*. Behaviour selection in his model consists of a competition among the agent's associated systems (behaviours). Hull proposed that the strength of a drive (therefore the **relevance** of the drive) responded to a multiplicative formula, namely: $drive = habit \times stimulus$. The strength of a *drive* results from the learnt *habit* and the *stimulus* strength, which depends on external and indeterminate stimuli, as expressed by $stimulus = f(internal,\ external, other\ stimuli)$. According to Hull, each action system has an associated drive, and the selection is the process of priming systems whose related drives exhibit the highest values.

  At the level of implementation, these principles of selection are insufficient, as they do not formally specify how to relate drives to behaviours, nor do they specify formal procedures for the combination of stimuli nor how to choose among several activities.

- Tinbergen (1953) introduced a hierarchical node-based architecture, for which the relevance of a behaviour is determined via competition between active nodes. These are nodes whose *innate releasing mechanisms* are activated by certain external stimuli. When these are active, activation levels can flow to the inferior level in the hierarchy until reaching the motor command level. A single action per level is released. The **combination of stimuli** is **multiplicative** at each level of the hierarchy.

- The model of Lorenz (1971) only addresses how to generate the strength of a drive. The

drive disinhibits the behaviour, and is modelled as a water reservoir. Internal stimuli are abstracted as water streams flowing into the reservoir, regulated by a set of parameters (k constants). External stimuli flow into the same reservoir (S_p). Hence, the influence and **combination of stimuli** in defining the strength of the drive are **additive**. The strength of the drive results from the addition of these stimuli minus the inhibition provoked by higher (cognitive) centres, modelled as a negative flow.

- The model of Baerends (1976) is hierarchical; lower nodes are controlled by other higher nodes in the hierarchy. The **combination of stimuli** is mostly left unspecified, although can be assumed as largely **additive**.

- Maes (1991) proposed a distributed, non-hierarchical model of action selection based on activation nodes, which represent the relevance for each action. Furthermore, Maes' model encodes all possible relationships among these nodes, whether the nodes are consummatory or appetitive. In doing this, their conflicts, goal-achieving relationships for some goals and goal-counteracting effects between nodes and goals, have been considered in the design. External stimuli and internal stimuli are evaluated in two steps at each node. The resulting **combination of stimuli** is **largely additive**, although their combined effect depends on the nodes activated at the previous step.

- Rosenblatt and Payton (1989) proposed a hierarchical, feed-forward network model in a similar manner to Baerends'. However, unlike his, the new architecture foresees the *combination of preferences*. A single node is active at each level and receives connections from the relevant external and internal stimuli. Since each node consists of a neural network, the **combination of stimuli** may respond to any sort of function (not necessarily additive or multiplicative) that the network may implement. Activation propagates down to the level of actions, where the node with the highest activation inhibits the others and disinhibits its related action.

- The turning point of these architectures have been Tyrrell's approach (Tyrrell, 1993) and Spier's architecture (Spier and McFarland, 1996). Spier's *drk* model introduces the possibility of having an internal physiology to close the perception-action loop. This architecture combines stimuli in a multiplicative manner ($drive \times environment's related cue \times cue regarding every tool$). Therefore, the execution of a behaviour in this model follows a multiplicative rule.

These architectures are based on the assumption that the combination of stimuli responds to an additive or multiplicative formula except Rosenblatt and Payton. Using formulae imposes serious restrictions, since these *formulae* cannot explain some behavioural responses (Tyrrell, 1993), pp. 172. Furthermore, I argue that a more ambitious explanation of behaviour selection

cannot only be based in the study of the combination of stimuli. Behaviour selection always occurs in a context. I therefore propose that any framework for behaviour selection consider the *principle of ecology* (Pfeifer, 1994). The agent (natural or artificial) is part of a niche, hence an explanation of the way stimuli combine should naturally arise if the relationship between the agent and its environment is well understood and faithfully modelled. Related to this is a recent formulation in neurophysiology and robotics, which suggests that the selection of behaviour works on the basis of the comparison of future expected reward for each behaviour (Rolls, 2003). Therefore, if the combination of stimuli relates the inner motivation to the intensity of each behaviour, reward does in turn relate the outer experiences to the inner motivations. Although using reinforcement learning is not novel in AI, there is not a single example of an ecological framework that integrates reinforcement learning so far. Nevertheless, I have considered it appropriate to mention two examples of frameworks that use reinforcement learning. These do not have an internal physiology, therefore, they do not ground their interactions within their environment. However, they still pursue the same goal of adapting to the environment.

- The model of Blumberg (1997) introduces a hierarchical structure, where behaviours are organised in groups within which a single behaviour is chosen via lateral inhibition. Winners at each group compete then at the next level of the hierarchy until there is a winner at the lowest level. This gains control of the agent's actuators and is therefore executed. Blumberg's assumptions with regard to the combination of stimuli are based on McFarland's thesis: "behavioural activation is determined by natural selection" (McFarland, 1993). The consequences for a synthetic approach lead towards empirical tuning methods: "the modern (synthetic) approach is to regard this question as an entirely empirical matter, not subject to any particular discipline" (Blumberg, 1997). This is explicitly formulated in a temporal difference framework (Sutton and Barto, 1998), where three sorts of combination of stimuli are modelled: multiplicative, additive and a combination of these.

- Another model of interest from the viewpoint of robotics was proposed in Gershenson (2001); González Perez et al. (2000). Their method for the combination of stimuli is calculate according to the following equation:

$$A_i^C = O_i^E * (\alpha + \sum_j Fa_{ij}^S * O_j^S) + O_i^D, \tag{2.1}$$

  where A_i^C is the certainty value with which will be created the solution element C_i in the proprio/extero/drive congruent level of the blackboard (analogous to the motivation to select a behaviour). α is the weight or importance attributable to the internal state, O_i^E is the internal signal, O_i^D is the signal created in the drive level, O_j^S are the external signals associated to the internal state O_i^E and Fa_{ij}^S are the coupling strengths of the elemental behaviours. This model mostly focuses on behaviour selection only.

- Likewise, Redgrave et al. (1999) presents a model of action selection inspired in physiological measurements of the basal ganglia, which calculates salience directly on the strength of the stimulus. The stimuli are directly related to the potentiality of execution of a behaviour. This is a model of behaviour selection inspired in physiological observations of the basal ganglia. This model does not perform learning.

Therefore, despite the use of reinforcement learning, these frameworks do not use reinforcement learning for instrumental learning in a strict sense, since decision making still occurs as a result of an additive or multiplicative combination of stimuli.

The framework I introduce can encompass any combination of stimuli that a feed-forward neural network reaches to calculate. This implies a wide range of possibilities to combine stimuli, only constraint by the physics of interaction of the agent with its surrounding environment. It is expected that patterns of combination of stimuli (closer to the multiplicative or to the additive rule) will arise at an experimental level depending on the environment while the agent learns to use its resources in a beneficial fashion.

In addition to the manner in which stimuli are combined, *persistence* has also been considered as a good feature to assess a behaviour selection system. This is described next.

2.2.2 Persistence

Relevant from an adaptive perspective is the notion of *persistence*: "to persist in a state, enterprise or undertaking in spite of counter influences, opposition or discouragement" (Webster Dictionary Online). In roboticist terms, this refers to extending the execution of a behaviour while reasonable; usually until the intended effect is obtained.

Persistence has been qualified as a requirement for intelligent behaviour selection (Tyrrell, 1993). However, the underlying mechanisms of persistence have not still been revealed. Some contemporary models (Blumberg, 1997), influenced by Tyrrell (1993) and by classical ethological models (Baerends, 1976; Lorenz, 1971), have introduced persistence in their design considerations, in an attempt to test its effect at an experimental level. In a complementary fashion, the model of Gurney et al. (1998) introduced persistence in a bottom-up fashion by linking the duration of the execution of a behaviour to the role of non-phasic dopamine (DA) among neurons in the vertebrate's' basal ganglia (non-phasic DA indirectly modulates the difficulty of initiating and maintaining the execution of a behaviour). Finally, the model proposed by Avila-García and Cañamero (2002) implements persistence from an ethological perspective in an artificial robot. Extensive experimentation with this model has demonstrated the relevance of including persistence in any robotic behaviour.

Based on this, persistence has been considered as an element to assess the quality of a model. This, together with the procedure to *combine stimuli*, are the two main issues intro-

duced in this section. As documented, these have been addressed in previous models in an analytical manner. However, unlike these, this thesis intends a synthetic approach. It introduces a hypothesis regarding the process of *grounding knowledge* from the environment (namely learning affordances) and utilises constraints derived from the adaptation methods used in nature for selecting behaviours. The hypothesis says that these principles constrain the fashion in which persistence and the combination of stimuli occur. Furthermore, that the interaction between the agent's surrounding environment and the agent's internal physiology determines the details along which this occurs.

The process of grounding knowledge from the environment consists of defining the semantics of the objects in its surrounding environment for a particular agent. Therefore, this can only be defined by the agent itself by relating the effect of a particular behaviour and sensory pattern with its internal physiology. Implicitly, the dynamics of interaction with the environment are directly linked to the dynamics of the agent's internal needs. Because of this reason, it has been considered appropriate to devote the next section to describe the agent's internal physiology and its dynamics.

2.3 Homeostasis and Internal Physiology

Homeostasis is directly related to motivation. Homeostasis was first studied by Claude Bernard during the 19th century (Bernard, 1878). However, the first formal definition is attributed to W. B. Cannon: "*The condition of a system when it is able to maintain its essential variables within limits acceptable to its own structure in the face of unexpected disturbances*" (Cannon, 1929). This notion is directly related to the process of *dynamic self-regulation* in Maturana's and Varela's ontology (Maturana and Varela, 1980).

Homeostasis was used by Hull (1943) to frame the concepts of drive and of homeostatic variable to describe internal bodily processes for the management and regulation of internal resources. Drives express the status of deficit or excess of the homeostatic variables, which can be compensated either via internal self-regulation or via external interaction (McFarland, 1990). The capacity of compensation in either way is related to the stability of the system (Ashby, 1965), hence survival depends on maintaining essential physiological variables within their range of viability. Ashby's notion of viability has inspired the notion of empirical viability indicators (Avila-García and Cañamero, 2002).

Although homeostasis has been traditionally classified as a purely internal process (Cannon, 1929), I argue that both processes of **internal self-regulation and external compensation can be viewed as part of the same regulatory mechanism**. In fact, the agent is endowed with a set of courses of action related to the agent's environment, each of which requires a certain set of elements in the environment in order to have some effect on the internal variables, hence

to close the regulatory cycle. For example, the presence of something edible is necessary to execute the behaviour *eat*. The behaviour execution has then a compensatory effect on the level of hunger of the agent.

There have been several architectures having implemented homeostasis (Avila-García and Cañamero, 2002; Cañamero, 1997). In particular, Canamero's architecture is the first implementation of the aforementioned regulatory cycle. This initiates by the homeostatic variables (e.g., nutrition, stamina) expressing an urge for compensation via a set of internal drives (hunger, tiredness). Compensation occurs via behaviour execution (external compensation), which requires interaction with some elements in the agent's surrounding environment in order to be successful. In particular, the behaviour executed has to exert a beneficial effect on the agent's internal drives expressing an urge for compensation (for example, if the external stimulus is hunger, the right behaviour to execute is eating, since this diminishes the internal drive and increases the level of nutrition). In these terms, an agent adapted to its environment will be able to appropriately select behaviours to maintain its physiological needs within the boundary of survival. However, the adaptiveness of an agent can only be demonstrated if the environment or the requirements of the agents vary with respect to one another. Its adaptiveness will then depend on its capacity to modify its behavioural patterns until this compensation is sufficient to regain the stability of its physiological needs satisfied for the given scenario. The procedure to this end is however not straightforward, since there are as many possibilities of change as criteria of correctness. Nevertheless, some criteria can be derived from *Ashby's notion of viability* (Ashby, 1965), which I have implicitly assumed as a necessary criterion that the agent has to respect to survive. Therefore, this can be used as an assessment criterion for adaptive processes based on whether they succeed in bringing the agent onto its physiological stability or not.

Each of the architectures mentioned in the previous section addresses behaviour selection as the main adaptive process, disregarding the contribution of perception as an adaptive process. However, I argue that, from an ecological viewpoint, perception should be considered as part of the homeostatic cycle as an adaptive process. Accordingly, the architecture I propose in this thesis considers perception as a dynamic process, where objects are viewed from the agent's perspective in a functional manner (I often refer to this as to "affordance based perception"). I argue that the agent can learn affordances by monitoring the internal physiological dynamics for every behavioural response of the agent. This also suggests that learning affordances is a fashion to ground knowledge from the environment (Harnad, 1990), since the meaning of any object with regard to every agent is different, depending on the own set of goals and on the assessment criterion. Related to this is the notion of semiotic triangle (Suonuuti, 1997), which identifies three related issues in perception: the sensation of the object itself, the meaning and the sign used to represent it. In these terms, solving the grounding problem would mean to

define these three notions as well as their interrelation. This argument will be further continued in section 2.6.

Ecology is a principle stressing the mutuality of the interactions between an animal and its niche (Gibson, 1966). Reorienting this for a synthetic approach is a complex task, which we propose to facilitate through a convenient description of an agent as a set of inter-related units: **its goals, its needs, its morphology, its behaviour repertoire and its perception**. Although this section solely focuses on **perception**, these concepts interrelate and more importantly, justify the framework and the assessment criterion used in *the learning process*, which is introduced in section 2.4.

2.4 Learning to Select Behaviours

This agent views the environment from a functional perspective, learning the affordances (functionalities) offered by every object. However, the problem of selecting among different courses of action is easily encountered when an object offers more than one affordance to the agent. Therefore, selecting behaviours and learning to do so in an adaptive manner is a necessary ability for an adaptive agent. This section reviews previous viewpoints while introduces the ecological perspective with which I have formulated my approach.

The notion of viability introduced in the previous section asserts that adaptation depends on the ability to maintain the internal physiology stable when there are changes in the environment or in the agent's physiology. A procedure to this end consists of modifying the behavioural patterns, which directly relates to the topic of *learning to select behaviours*. There is a variety of methods used to this end in robotics. This section reviews them before describing the procedure used in the model introduced by this thesis. Two learning methods have been broadly addressed in robotics: learning by imitation and learning by reinforcement. These are described next.

Learning by imitation consists of transferring procedures to execute actions from demonstrators to learners, either directly by observation or via social facilitation. For the former case the information is passively demonstrated, for the latter this is actively performed. The procedure may vary depending on the knowledge to be acquired, however, as an example for the case of motor commands, it can be decomposed to the set of sub-processes introduced below. It requires the learner to decode visual information from the movement of the limbs of the demonstrator and furthermore, to correctly associate the demonstrator's limbs with its own ones in order to infer the motor commands that the demonstrator was using. This sort of learning has been observed to be part of the developmental abilities of higher primates (Heyes and Galef, 1996). In support of this, recent neurophysiological measurements suggest that the repertoire of motor commands is encoded in the motor cortex of some primates and that several types of neurons co-interact, mainly the ones encoding the motor command itself (motor-neurons),

and those capable of matching the observation of the execution of a motor command to its own repertoire of commands—*mirror neurons*—(Rizzolatti et al., 2000). This method, joint with inspiration from ecological psychology (Gibson, 1966), has reached the level of implementation in robotics for the case of motor commands (Metta and Fitzpatrick, 2002; Demiris and Hayes, 2002; Maistros and Hayes, 2001; Billard and Mataric, 2001). Other sensory modalities have also been addressed in infant psychology (Kessen et al., 1969) throughout the last decades, although not until recently have they reached domains of artificial intelligence and robotics (Westermann and Miranda, 2002; Kohler et al., 2002). Mirror neurons have been located in higher vertebrates, such as macaques or humans only. These are hypothesised to match perceptual information with an appropriate set of motor commands[2]. However, learning by imitation has also some limitations, since two or more individuals endowed with complex hardware are required. Learning by imitation is supported by medical studies on language and motor learning disorders. Furthermore, understanding this would save tedious programming hours for roboticists.

However, this thesis focuses on individual learning principles, based on the mutual relationship between the agent and its environment, namely on the responses that the agent receives by interacting with the environment. This relates to a whole family of learning algorithms, usually referred to as *Learning by Reinforcement*. For every particular application in reinforcement learning, independently from the framework, the goal is reached via sequencing a set of courses of action in a manner that reward is maximised throughout time. Similarly, in the model described in this thesis, the goal is the minimal deficit and physiological stability, which are reached via learning to sequence actions in an appropriate manner. These assumptions have been selected for two main reasons. Firstly, they partly fit some biological data; Schultz's experiments (Schultz, 1998) support the hypothesis that reward —or its absence— may play a role in strengthening or weakening the neural representations of instrumental relationships between stimuli and actions, and between actions and their effects. Secondly, because this is is a natural extension of an affordance based framework to sequence courses of action.

A Brief Review of Reinforcement Learning in Robotics My interest is finding a plausible explanation to the problem of adaptation within the boundaries set by the reinforcement learning framework. From this perspective, robotics is a broad field of application of the aforementioned techniques. One of the first —most renamed— examples of this cooperation was Lin (1993), who proposed the use of reinforcement learning as a possible pathway for overcoming limitations of reactive architectures in terms of planning and adaptive action selection.

Towards this direction, several authors have used different formulations of reinforcement learning algorithms to address learning in the robotic context. In *Navigation*, Toussaint (2003)

[2]The term motor is used in a broad sense, meaning any action related to the sensory modality responsible for its activation, e.g., to see a chair and to sit.

has recently provided a nice approach to adaptive solutions for creating models of the world to be used for navigation on the basis of reward functions. Another interesting solution combining reinforcement learning and heuristics has been proposed by Konidaris (2003). A more abstract approach was followed by Blumberg (1997), who proposed the use of temporal difference —a particular reinforcement learning update rule— within a hierarchical architecture for guiding the patterns of behaviour of a believable artificial creature. An example of emotional architecture to bias behaviour selection based on reinforcement learning was proposed by Gadanho (1998). From a more theoretical perspective, Humphrys (1997) proposed a possible solution to extend reinforcement learning to be able to scale up to high dimensional spaces with the use of W-learning, a sort of hierarchical set of small reinforcement learning algorithms.

Although in theory reinforcement learning could be applied to any sort of Markovian domain, most of the aforementioned applications are limited to discrete domains. A thorough review of these methods is Sutton and Barto's book (Sutton and Barto, 1998). Further application to continuous domains has been mostly approached during recent years from a theoretical perspective (Doya, 1999). These studies are progressively reaching one of the most challenging fields in AI, that of humanoid robotics. Beyond being a pure application, it is serving as a benchmark of development of interdisciplinary solutions for two of the most critical domains of reinforcement learning: continuous and high dimensional spaces (Peters, 2003).

The set of applications of reinforcement learning is currently very large, however, it is still possible to provide a standard formulation for a generic framework for reinforcement learning. This is shown in figure 2.1.

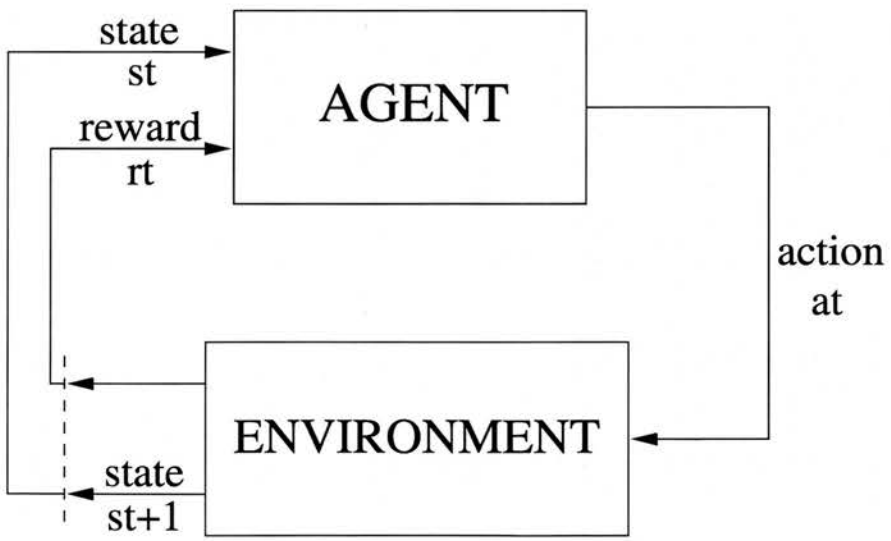


Figure 2.1: Framework of a Generic Reinforcement Learning Problem. a_t is the action at time t, s_t is the state at time t, s_{t+1} is the state at time t+1.

The elements of the algorithm are: a set of Markovian states s_t describing the state of the

agent, the action a_t (to be chosen among the agent's repertoire), to move from one state to another and the reward r_t function, which quantifies the effect of the execution of each action on the environment and the environment itself. The definition of *reward* is fundamental and specific to each problem, since this is the definition of good- and badness that will reflect on the desirability of one behaviour over another. From a biological perspective, reward has been often referred to as *sense of valence* (Ackley and Littman, 1991), and has been hypothesised to be a mechanism to increase the individual's chance of survival (Damoulas, 2004). The second element related to the reward is the definition of *state*. It must be considered that reward is delivered when following a transition among states, therefore the states must sufficiently characterise the environment to make these transitions meaningful. Otherwise, some states may not be distinguishable, which would make related decisions uncertain. This could lead to a dramatic decrease in performance (Crook and Hayes (2003) — two states look alike though they are in different locations in the state space). The difficulty of addressing this is beyond the reach of this thesis. To overcome this, environments have been engineered to be easily distinguishable among similar states.

These elements introduce the framework to view learning as an extension to the simple change of activity, where reward may be used to modify responses to stimuli and to adapt to changes in the environment. To be able to do this, I am implicitly assuming that reward is related to effect and that every action has an effect relating to the animal's physiology. For example, eating increases the level of glucose in blood, jogging (if exercised enough) increases stamina. Therefore, I view reward as a measure of the effect from the perspective of the agent. This depends on the internal, physiological and affective state and on the surrounding environment. Loosely speaking, reward could be viewed as the **feeling** experienced by the agent due to the physiological effect. The definition of reward is further explained in the chapters 4 and 5. This argument is further grounded on principles of neuroscience introduced in the next section.

2.5 Neuroscience Background

A possible explanation to the interest raised by reinforcement learning is not due to the robotic applications, but to the physiological measurements performed by Wolfram Schultz of the Ventral Tegmental Area (VTA) and in the Substantia Nigra Pars Compacta (SN_c) (Suri and Schultz, 1998; Schultz et al., 1997). These suggest that neural activity in the output nuclei of the basal ganglia signal the error in the prediction of reward given a certain stimulus (Fiorillo et al., 2003; Schultz, 1998). Hence, the hypothesis is that one of the roles of the basal ganglia is the learning of stimulus-response relationships. The experiments to test this consisted of presenting a stimulus followed by a reward to a macaque monkey. Dopamine (DA) from the SN_c and VTA signals the novelty of this reward for the first trials of the experiment and decreases at later

trials. The hypothesis is that after some trials, the system learns to predict the rewarding event given the stimulus. A computational model including these effects was introduced by Dayan and Montague (Schultz et al., 1997) and is further described in section 2.5.2.

This hypothesis for the role of the basal ganglia solely includes learning and is unrelated to action selection. Nevertheless, there is a second hypothesis arguing the possibility of the basal ganglia being a centralised action selector (Redgrave et al., 1999). This is based on the anatomical evidence that the cortex has a large projection to cells in the input nuclei of the basal ganglia (the striatum). Furthermore, the output nuclei, the SN_r and the Globus Pallidus Internal Segment GP_i, project back to the thalamus and to the cortex. These closed-loop connections suggest that cells in the striatum receive much information about the situation that the animal is experiencing. This also hints that the projections between the striatum and the GP_i/SN_r are parallel pathways, each devoted to a behaviour, which is disinhibited by the activity of the cells in the striatum.

These two hypotheses have been so far irreconcilable. However, the learning hypothesis of Schultz can be easily extended to action selection if Pavlovian learning is viewed as a particular case of instrumental learning, where the delivery of reward is mediated by the absence of action. This approach has been followed by several authors in machine learning and robotics (Sutton and Barto, 1998), however, only recently has it received some attention by neuroscientists (McClure et al., 2003). This suggests that there is a simpler solution to the problem of action-selection than that proposed by neuro-scientifical hypothesis. This will be tested in the experimental chapters.

The following sections introduce both models of the basal ganglia (Redgrave's and Dayan's —inspired by Schultz's ideas) in order to frame and introduce an extended version of the actor-critic, which is introduced thereafter.

2.5.1 Redgrave's Model for Action Selection

A biologically inspired model for action selection has been proposed by Redgrave et al. (1999) and (Gurney et al., 1998). This hypothesises that the main role of the basal ganglia is to be a centralised action selector. Despite conflicting with the model of Schultz et al. (1997), it has demonstrated its performance in a robotic architecture (Gurney et al., 2001b,a). The model is shown in figure 2.2 at a functional level.

The structure of the basal ganglia is considered as an input/output system, which receives afferent projections from the cortex to the striatum and whose output nuclei, the globus pallidus and the substantia nigra, project to the cortex and to the thalamus (the connections have been omitted from figure 2.2 to facilitate its understanding).

The action selection process in Redgrave's model is divided into two different sub-processes: **selection** and **control**. The model also distinguishes between D1 and D2 dopamine receptors

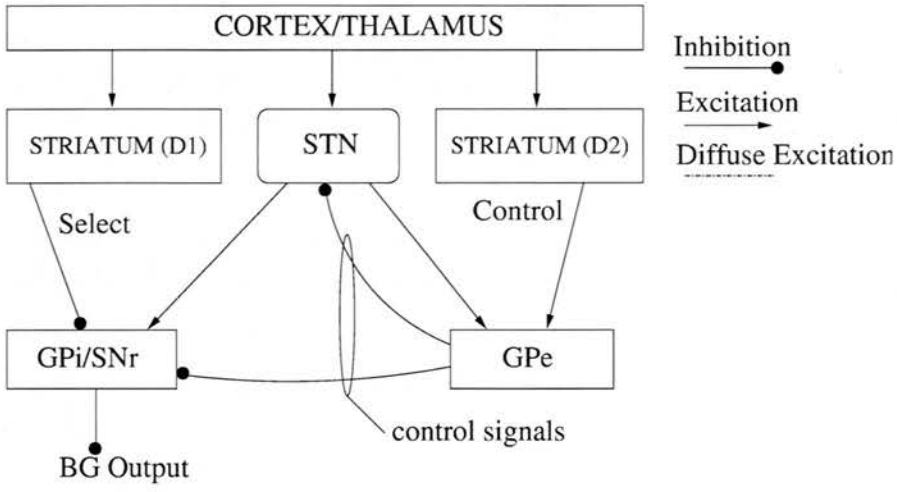


Figure 2.2: Redgrave's Model of Action Selection. STN=Sub-Thalamic Nucleus, GP_e=Globus Pallidus External Segment, SNr=Substantia Nigra Pars Reticulata, GP_i=Globus Pallidus Internal Segment, BG=Basal Ganglia.

(D1 receptors affect an excitatory response, D2 inhibitory) in the striatum.

The striatum is composed of several neural centres each of which encodes the **salience** of a single action, based on a set of signals from the cortex and the thalamus. Via lateral inhibition, these centres compete with one another for activation, resulting in a reduced number of active centres. It should be noted that the physiological evidence for such lateral interactions is still weak, although this is not strictly necessary to perform action selection. The neural centres in the striatum project via inhibitory connections to the output nuclei, which are released only when the striatal neural centres are active, i.e., acting in a disinhibitory manner. This sub-process involves the so-called direct pathway in the basal ganglia (Prescott, 2001). Redgrave et al. also introduce a secondary pathway, whose hypothesised role is to refine the selections suggested by the primary pathway. The secondary pathway involves the Globus Pallidus External Segment (GP_e) and the cells in the striatum with D2 receptors. They suggest that this pathway automatically scales excitatory outputs. The GP_e sends out control signals that project to the Sub-Thalamic Nucleus (STN). Negative feedback to the STN scales the outputs of the active channels. A second hypothesis concerns the synergistic action of dopamine in both the *control* and *selection* pathways. The hypothesised role of dopamine (DA) in this mechanism is to *regulate the ease of selection*. Hence, an increase of DA should provoke promiscuous selection, and its absence a high level of inhibition (consistently with final stages of Parkinsons disease).

If any of the neural centres in the striatum are active, this introduces the possibility of them being further inhibited, from the Sub-Thalamic Nucleus (STN) to GP_e and from this to the output nuclei in the SN_r. Hence, this would facilitate or complicate selection, acting as an

activation threshold of the output nuclei.

This model is mostly engineering based. It includes the notions of *clean switching* between behaviours and *persistency*; which are desirable for selection in the basal ganglia. Nevertheless, their hypothesised dependence on DA remains uncertain.

The model of Redgrave was embedded into a Khepera robot for testing purposes. This calculates salience the sum of relevant variables: *perceptual*, *motivational*, *positive feedback* and *efferent copy*. The first two are, respectively: affordances present in the agent's surrounding environment and the internal state of the agent (its drives). The positive feedback has been proposed as the ganglia-thalamo-cortical loops which may act to provide a positive feedback pathway (derived from the efferent copy) that can maintain an appropriate level of salience of a behaviour (hence forcing persistence). Each behaviour may be able to generate a "busy" signal that contributes to its own saliency.

In conclusion, this model introduces the hypothesis that the Basal Ganglia working as a behaviour selector and considers the level of non-phasic dopamine (DA) as a threshold controlling the selection of the behaviours. However, it considers neither learning nor the role of phasic dopamine in its model.

2.5.2 Dayan and Montague's Model for Learning

Unlike the previous model, the hypothesis of (Schultz et al., 1993) was made explicit via Dayan and Montague's model. Their most relevant contribution has been to show that Schultz's hypothesis of the Temporal Difference (TD) algorithm as a possible explanation of the learning phenomenon was consistent. Both the hypothesis and the measurements have solely addressed learning, although as it will be argued next, this naturally relates to action selection.

Schultz's measurements in monkeys suggest that phasic dopamine signals effective reinforcement that mediates Pavlovian learning in the vertebrate's brain. Schultz's experiments measured DA activity from cells in the Ventral Tegmental Area (VTA) and/or in the Substantia Nigra Pars Compacta (SNc). The procedure consisted of exposing the monkey to a stimulus, followed by a reward some seconds later. The measurements of the DA cells show a significant increment of their spiking frequency the first few times the reward is given. This decreases gradually after repeated trials until reaching its stationary level.

Schultz's explanation of this decreasing phenomenon is bound to the hypothesis that some part of the basal ganglia is implementing a TD learning algorithm (Sutton and Barto, 1981). In this light, the basal ganglia is viewed as an Input/Output device. The input nucleus to the system is the striatum, and the output nuclei are the Globus Pallidus Internal and External segment (GP_i and GP_e) and the Substantia Nigra Pars Reticulata (SN_r). These are represented in figure 2.3 along with the Thalamus. Anatomically, it has been observed that cells in the striatum receive strong projections from the thalamus and the cortex, and that dopamine projections

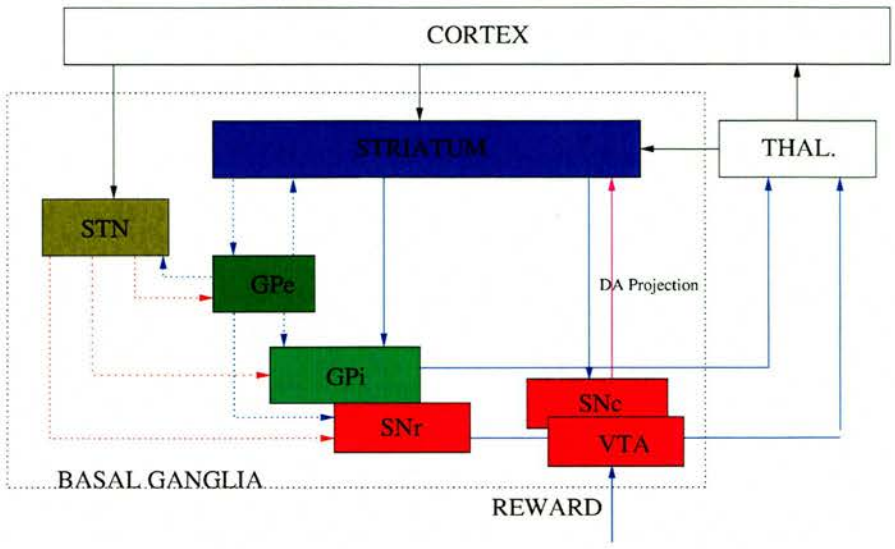


Figure 2.3: Basal Ganglia Functional Representation. The abbreviations are: THAL for Thalamus, VTA for Ventral Tegmental Area, STN for Subthalamic Nucleus, SN_r for Substantia Nigra pars reticulata, SN_c for Substantia Nigra pars compacta. GP_i for Globus Pallidus Internal Segment, GP_e for Globus Pallidus External Segment, DA for Dopamine. The different colours of the arrows and boxes have been only used for clarification purposes.

from the SNc and VTA to the striatum are hypothesised to behave as an effective reinforcer, i.e., signalling the error in the prediction of reward following that stimulus.

The Temporal Difference (TD) algorithm was proposed by Sutton and Barto in the early 80's (Sutton and Barto, 1981) as an instrumental learning algorithm, and was further formalised by Dayan and Montague as a model of the Basal Ganglia (Schultz et al., 1997).

TD assumes that learning is aimed at maximising future rewards, hence at maximising the following function:

$$V(s_t) = E[r(s_t) + \gamma r(s_{t+1}) + \gamma^2 r(s_{t+2}) + ...] \tag{2.2}$$

$V(s_t)$ in equation 2.5.2 is the state-value function in the reinforcement learning context, and equals the addition of the rewards due to transitions between states, from time t until the goal (stability) is reached. γ is the discount rate (ranges between 0 and 1). In this model, the TD algorithm also assumes that the current state, defined by the cortical and thalamic projections, only depends on the previous state (Markovian property). It uses the quantity

$$\delta(s_{t-1}) = r(s_{t-1}) + \gamma V_t(s_t) - V_t(s_{t-1}) \tag{2.3}$$

as the effective reinforcement at time t; where $\delta(s_{t-1})$ is the effective reinforcement for state s_{t-1}, $r(s_{t-1})$ the reward at time $t-1$ and $V_t(s_t)$ the effective reinforcement at time t for state s_t. Therefore the update rule relates $r(s_{t-1})$, the reward due to the behaviour executed at a

time $t-1$ to its predictions of cumulative reward before and after that behaviour has been executed, $V(s_{t-1})$ and $V(s_t)$. The corrections occur for individual transitions; however, the goal of the learning process is global; physiological stability. In the long run, when the appropriate transitions to the goal have been learnt, the return will be predicted correctly and the value of δ will be 0. Furthermore, a similar update rule affects the critic itself after every transition, the value of the V function (the critic) for the previous state s_{t-1} is updated according to this equation

$$\Delta V_t(s_{t-1}) = \alpha \times \delta(s_{t-1}), \tag{2.4}$$

whereby α is the learning rate. Concluding, the hypothesis relating both the algorithm and the observation suggests that cells in the patch of the striatum embody the V (state value) function, and that the dopamine projection signals the δ, the error in the prediction of reward. This would explain the silence of the dopamine neurons after several: the prediction of reward is then correct, therefore δ is 0.

This model introduces the role of learning, i.e., of learning to associate stimuli to rewarding events. However, this has been, at an experimental level only measured in Pavlovian contingencies. In this respect, I suggest that there is sufficient evidence to extend this model to instrumental learning and to extend the hypothesis of DA as a predictor for learning general behavioural patterns. This is the role of the actor-critic (Sutton and Barto, 1981) introduced in the next section.

2.5.3 Justification of an Actor-Critic

The last two sections have described two models of the basal ganglia. These have been built based on the hypotheses that the main role of the basal ganglia is either learning to relate stimuli to rewards or action selection. However, I argue that it is quite likely that both roles (among others) are been concurrently performed in the basal ganglia, and that both functions interact with one another. In the path towards a more general model, I suggest to integrate both hypotheses in a single model that learns to select behaviours: the actor-critic model.

The actor-critic is a reinforcement learning algorithm embedding both learning and behaviour selection in two modules of the same architecture. The actor selects behaviours and modifies its policy on the basis of the signal delivered by the critic, which corrects the predictions of future return for that behaviour. Therefore, this algorithm would already encompass both Schultz's and Redgrave's hypotheses. This is not thoroughly supported experimentally, but has already been suggested as a future working hypothesis (Houk et al., 1995). Both hypotheses are described next, together with their relationship to the actor-critic.

With regard to *behaviour selection*, Redgrave et al. (1999) suggested that several neural pathways projecting from the thalamus and the cortex onto striatal cells provide information to select among the output pathways projecting onto the basal ganglia's output nuclei (to disinhibit

a single behaviour). The striatum is therefore processing every piece of incoming information and calculating the intensity of every outgoing pathway. Behaviour selection is then conceived as a competition between several output pathways. This competition is possible if I assume the notion of reward according to Rolls (2003). Rolls argues that the selection among different courses of action is driven by the maximisation of future reward. If this is correct, it would be sensible that Redgrave's outgoing pathways were measured in terms of future reward. The behaviour predicted to return the highest reward would be selected to be executed. Therefore, I am hypothesising a direct correspondence between Redgrave's common currency and Rolls' reward. Based on this, I assume that the actor (within the actor-critic) is assuming the role of selecting a behaviour over another.

The experiments performed by Schultz et al. (1993); Schultz (1998) with regard to *learning* suggest that the dopamine projection from the SN_c and VTA to striatal cells signals the error of the prediction of reward when learning to relate a stimulus to a response in a Pavlovian contingency. I argue that this same reward is acting as an internal assessment to correct the predictions of future reward and for biasing the selection of one behaviour over another. Therefore, by using the actor-critic, I am extending Schultz' hypothesis to instrumental contingencies as well. Contingencies for which the delivery of reward is mediated by the execution of a behaviour. In this light, I argue that reward is modifying the neural structures predicting reward for every behaviour.

I intend to use this framework to study the manner in which reward-based mechanisms contribute to adaptation and to answer related questions: **how to combine stimuli** and **how to integrate appetitive and consummatory behaviours**. I argue that the answers to these questions may emerge if behaviour selection and learning are considered as two concurrent processes integrated in a single ecological framework as the actor-critic. In addition to this, I am also concerned about the way agents learn to perceive their environment as a function of their goals and motivations and how they learn to make decisions based on this information. In a way, these open questions require an understanding of the topics to be addressed in this thesis. Firstly, the understanding of the dynamics of the animals' **internal physiology**. Secondly how to learn to perceive the environment, **how to ground knowledge from the environment** based on the feedback provided. Finally, **the processes underlying the process of behaviour selection**. Concluding, the aforementioned hypotheses suggest that the actor-critic is the biological mechanism integrating integrate both sensory and motor information to select among future courses of action. It is therefore interesting to build an artificial model that would concurrently consider ecological perception, behaviour selection and learning in a natural manner. This concurrency must be considered since adaptation cannot occur in the absence of a single dynamics coordinating resource management, behaviour selection, learning and adaptive perception.

Lastly, I also pursue the *extraction of biologically plausible conclusions about learning in*

perception and in behaviour selection. The remainder of this review addresses a description of the necessary elements to build an ecological perception system, and to integrate this in the actor-critic. The final section introduces the developmental approach followed throughout this thesis.

2.6 Ecological Perception: Gibson's Affordances

The previous sections have presented the principles relating learning and behaviour as interactive processes in an ecological context. In a complementary fashion, this section situates the notions of ecological perception and of affordance and explains the way in which these intervene for adaptation.

As mentioned in section 2.4, transitions for the actor-critic, as for any other reinforcement learning algorithm, occur between two states in a Markovian space. Therefore, part of the problem when defining the framework of the actor-critic is defining the states themselves. As a general rule, as for any other reinforcement learning algorithm, the states will have to provide sufficient information to appropriately make a decision. In this case, since the selection of behaviour occurs in an environment inspired after biology, it is most sensible that the state includes information from the agent's outer environment and from its own physiology. The part that depends on the agent's physiology is described in section 2.3. In a complementary fashion, this section focuses on describing a novel way to view the environment: a *functional view based on object affordances*. The state of the actor-critic is based on this. To commence the description of this ecological view, I have considered appropriate to review some historical concepts on traditional psychology. The following paragraphs are only intended to explain the chain of events leading to the concept of affordance from a historical perspective. However, the reader may decide to skip reading until the end of this section without losing the current line of thought.

Gibson's ecological perception theory originated as a counter-position to dualism, which was based on the assumption of a physical dimension, separated from a phenomenological dimension of nature. Along these lines, Gestalt theorists maintained that *objects in their phenomenological dimension tell us what to do with them*; according to Koffka, they have *demand character*, a concept further shaped by Lewis as *Aufforderungscharacter* and translated as *invitation character* by Brown (1929) and as *valence* by Adams (1931). In philosophical terms, the concept of valence is related to the phenomenological dimension of an object and it is disjoint from its physical dimension. Furthermore, its value is bestowed by a need of the observer. Thus, only when she or he needed to eat, did the food have its demand character.

The term *affordance* in the Gibsonian ontology derives from the aforementioned concepts. However, unlike these, the formulation of the affordance is based on the principle of ecology

and denies the existence of a phenomenological dimension. Citing Gibson: *"The object offers what it does, because it is what it is. Duality does not exist, the physical object (the only one), possesses meaning and value to begin with"*. Furthermore, this value is independent of the motivational state of the agent (the level of hunger, the level of tiredness, etc. of an animal).

Gibson's ecological perception opposed contemporary cognitivist views. Cognitivists argued that perception is a process integrating object features, which are perceived separately. Instead, Gibson asserted that affordances are specified in the structure of the ambient light, hence by the sensory signals perceived by an animal. Therefore, the perception of the object can be viewed (and simplified) in terms of *direct perception*, i.e., the meaning of the object is directly perceived in terms of the behaviours and actions offered. This is individual knowledge, valid only in the context of an individual observer.

Reformulating this in synthetic terms, to learn affordances consists of establishing the value of sensory information with regard to the agent's internal goals. This is equivalent to *grounding the meaning of sensory cues with regard to the agent's internal goals* (Steels, 1994) —or as I prefer to say, with regard to the agent's internal physiology. I suggest that this learning capacity offers the possibility of building agents capable of adapting to a variety of environments. This is the major principle guiding the design of the architecture introduced in the next chapters.

2.6.1 Perception of Function

Gibson's affordances relate a set of sensory cues to the potentiality of performing a behaviour. For example, an affordance of a chair or of a table is to support. This is a fashion to define objects.

Gibson (1966) postulated ecological perception on the grounds that an animal and its environment are part of a single entity. Therefore, animals (and plants) have arisen via mutual interaction with their environment. This suggests that every animal is probably bound to live in its own niche of creation. For example, dolphins are to some parts of the sea as humans are to some regions of the earth. In general, animals are prepared to manipulate a restricted environment in an intelligent manner and are sensitive to aspects of the environment useful to them. For example, the view of a prey for a fox is fundamental, but the view of a bottle of whisky might not be. Hence, perception seems to be functionally biased according to species, according to their internal needs. Gibson's theory of affordances agrees with this principle.

Two of Gibson's most relevant forerunners are von Uexküll (1921) and Thorndike (1911). Von Uexküll transferred the viewpoint to study nature from the observer to the animal: *"And due to the fact everything natural for us disappears: the whole nature, the earth, the sky, the stars, all and everyone and each of the things that surround us, and everything which remains is only the effect of elements from the world having some influence in the structure of the*

animal."[3]

In this respect, his *Funktionskreis*[4] defines a functional relationship between the agent and the environment: *"Each stimulus coming from the same attribute*[5] *will next be transformed into excitations of different nerves that meet together again in the centre, in a property-net meaning an action primitive."*[6] Furthermore: *"For the case of large animals, each property-net corresponds to a nervous effect-net, from which some paths lead out to groups of muscles that constitute an action primitive."*[7] Therefore, the *Funktionskreis* is a loop that relates sensation and action. *The object in its double feature set, as property- and effect-support, possesses its own structure binding together this double property* (von Uexküll, 1921).

Similarly, Thorndike (1911) argued that stimuli are structured into families, in which any situation was associated with a hierarchical set of responses. The response taken at any point in time was at the top of the currently applicable habit family, and is modified through the laws of exercise and effect (so there is promotion of responses that are successful and weakening of responses that fail).

Although not clearly detached from dualism, both Thorndike and Von Uexküll could be classified as non-cognitivists who argue in favour of a functional view of the animal-environment relationship as part of the same whole. Although there is no bibliographical evidence (to the best of the author's knowledge) that Gibson knew Von Uexküll's work, the concept of affordance is a natural continuation of the perception postulated by Von Uexküll as *Funktionskreis*. Furthermore, both found behavioural support in Thorndike's hypothesis of action priming and selection.

2.6.2 Affordances

Gibson's ecological perception introduces the view of the agent and the environment as part of a single entity, whose perception is not based on identification, but on differentiation. Hence, learning to differentiate is learning to perceive. Details of objects will reveal themselves after further interaction. Furthermore, the significance of an object with respect to an agent arises from this interaction and can be considered invariant with respect to that particular agent.

Gibson further extended this definition through the concept of *affordance*. The agent was modelled based on feedback and on sensory signals from the environment and its memories as

[3]Damit veschwindet alles, was für uns als selbstvertändlich gilt: die ganze Natur, die Erde, der Himmel, die Sterne, ja alle Gegenstände, die uns umgeben. Es bleiben nur noch jene Einwirkungen als Weltfaktoren übrig, die dem Bauplan entsprechend auf Tier einen Einfluss ausüben, von Uexküll (1921)

[4]Translated by D.L. Mackinnon as function-circle (Von Uexkull, 1926).

[5]It remains however unclear whether he refers to features, thus Merkmal in German means 'attribute' but also counts 'feature' as a possible translation.

[6]Alle von einem Merkmal stammenden Reize werden zunächst in Erregungen verschiedener Nerven verwandelt, die sich im Zentrum in einem nervösen Merknetz zusammenfinden und dadurch die Einheit des Merkmals schaffen.

[7]Jedem nervösen Merknetz entspricht bei höheren Tieren ein ebenfalls nervöses Wirknetz, von dem die Bahnen ausgehen, welche bestimmte Muskelgruppen zu einer einheitlichen Handlung zusammenfassen.

centres, each of which resonates to different sensory cues. Hence, the invariance of perception depends on the way these centres are developed to allow the invariants in the input to be differentiated at a neural level. The registering of invariants is something that all nervous systems are geared to do, even those in the simplest animals. When the sensory signals related to certain objects are perceived (shape, size, colour, texture, composition, motion, animation and position relative to other objects), the observer can detect their affordances. Their affordances depends on these. These are then classified into categories and subcategories according to their degree of similarity.

Gibson's affordances can be viewed as a concrete case of von Uexküll's *Funktionskreis*. In this respect, it can be argued that Rizzolatti's research (Rizzolatti et al., 2000) brings them together by explaining their underlying neural support. Some nuclei in the brain are active when certain actions or parts of objects are perceived (hence, resonate to the stored information). This allows agents to perform actions by self-observation and by mirroring other individuals. This implies that perception and action are closely coupled and that direct perception of objects and action priming are related by the affordance.

The concept of affordance is used in the literature in a very abstract manner. I have therefore considered it appropriate to conclude this introduction on ecological psychology by introducing a set of citations to frame my own definition, which is introduced at the end of section 2.6.3.

- According to Gibson (1986), from a general point of view *the basic affordances are perceivable and are usually perceivable directly, without an excessive amount of learning. The basic properties of the environment that make an affordance are specified in the structure of ambient light, and hence affordance itself is specified in ambient light. Moreover, an invariant that is commensurate with the body of the observer himself is more easily picked up than one non commensurate with his/her body*

- Affordances are properties taken with reference to the observer. They are neither physical nor phenomenal. Only physical objects exist.

- The notion of **invariants** that are related at one extreme to the **motives and needs of an observer** and at the other extreme to the **substances and surfaces of a world** provides a new approach to psychology.

In conclusion, the agent does not make sense without the environment, whose perception and abilities have been developed to match its need to survive in that particular environment. Hence, if we wish to build agents adaptable to different environments, they must be able to learn affordances in order to perform actions to satisfy their internal needs.

2.6.3 Uses of the Term Affordance

The concept of affordance has only found certain acceptance in neuroscience and psychology. However, a series of definitions can be met in the literature. This section lists a series of definitions and models using affordances I have considered most salient. Finally, it also introduces my own definition of affordances, which I will use next to build the model.

- In robotics and AI, its main field of application has been imitation, where an affordance has been defined as *the perceived or actual properties of how something may be used by the agent* (Nehaniv and Dautenhahn, 1998). Affordances are a bridge relating the agent's environment to its behaviours. Furthermore, Nehaniv proposes them as the unit of perception which could be used to facilitate imitative processes between agents of different morphology.

- In psychology Cooper and Glasspool (2002) explicitly built a model using affordances as attentional mechanisms that filter the amount of information which can be perceived from the environment. Their definition is however not entirely Gibsonian, since this is based on symbolic features perceived from objects, defined and engineered by the designer.

- In neuropsychology Fagg and Arbib (1998) postulated a model for the extraction of Gibsonian affordances. This is in the context of navigation and manipulation tasks. According to them, the term affordance is used to mean the link between the visual cues to parameters relevant for motor interaction.

- In Human-Computer-Interaction (HCI) St. Amant (1999) emphasises the situated nature of human behaviour. HCI designers view affordances as the perceived properties of a software artifact that indicates how it can be used (Baecker et al., 1995). St. Amant also defined them as: *"a mechanism that allows or facilitates the execution of some operator. More specifically, an affordance preserves the conditions necessary for the successful completion of the operator by reducing the execution cost of other appropriate operators or by increasing the execution cost of inappropriate operators."*

I argue that the difficulty to define the concept of affordance lies on its intuitive character. However, I argue that St. Amant (1999) introduced the most comprehensive definition of affordance, including four separate dimensions for the concept: first, Gibsonian affordances (relationships or properties of relationships), second our perception of several properties, the surfaces, distances, areas, textures, relationships between parts (on this fact relies the design of ecological HCI), third the mental interpretation derived from perception, fourth the act of performing an action itself.

Along with Gibson's view, I argue that *affordance has to be viewed as a relational concept*, established between an agent and its environment, since their meaning is restricted to that framework. I define *affordance* as *the potentiality offered by a set of sensory cues to a certain agent of performing a behaviour*. Therefore, affordances can only be defined from the perspective of the agent, and are grounded on the dynamics relating the agent's physiology to the environment.

The next section introduces the neuroscience data hypothesised to embody affordances. These data have been often referred to in models of learning by imitation (Demiris and Hayes, 2002), based on the mirror neuron hypothesis, also introduced in the next chapter. I have deliberately chosen the FARS model as the most complete model encoding animal affordances, since this is consistent with the parts of the brain also hypothesised to embody affordances.

2.7 Ecological Principles in Neuroscience

The structure of an animal is a result of the interaction with its environment. Hence, the tasks an animal may be able to perform are closely related to its environment. Furthermore, an animal can be viewed as an active entity performing some function related to the balance in its environment (thus its reason of existence); when the balance is impaired, the species adapts or extinguishes.

Hence, the internal resources of the animal are related to the maintenance of its internal energy and of the balance of its internal structure within the boundaries that enable life, so that individuals can transmit their genes to the individuals of the next generations for continuing their balancing function with respect to their ecosystem (Ashby, 1965).

Some recent neuroscience studies have addressed the principles underlying ecological perception and their possible relation to the inverse mechanisms for action generation (Oztop and Arbib, 2002; Rizzolatti et al., 2000; Guazzelli et al., 1998). The use of these principles for a robotic synthetic approach is nonetheless insufficient, since these principles do not integrate the environment and the agent's internal physiology in a dynamic defined by themselves. To this end, there have also been several elements relating to action selection, which will be introduced in the next section.

2.7.1 The FARS Model

The FARS (Fagg-Arbib-Rizzolatti-Sakata) model (Fagg and Arbib, 1998) introduces an implementation of the principles underlying perception and the generation of action responses. It is based on experimental studies in macaque electrophysiology. This model uses the term *affordance* to refer to a set of motor commands relating to a set of sensory cues.

The schema of the FARS model is introduced in figure 2.4, and mostly involves areas AIP

and F5 in the macaque brain. AIP (Anterior Intraparietal Sulcus) is located in the parietal cortex. It receives projections from the visual cortex and is hypothesised to extract the relevant set of cues to restrict the set of grasping actions encoded by the F5 area. Neurophysiological recordings in this area suggest that some neurons encode the set of motor commands executed by the monkey when grasping an object. Most importantly, some of these cells show the same activation pattern when the related motor sequence is *observed* by a demonstrator. Rizzolatti hypothesised that this suggests that these cells are acting as a bridge between perception and action, enabling the macaque to learn by imitation.

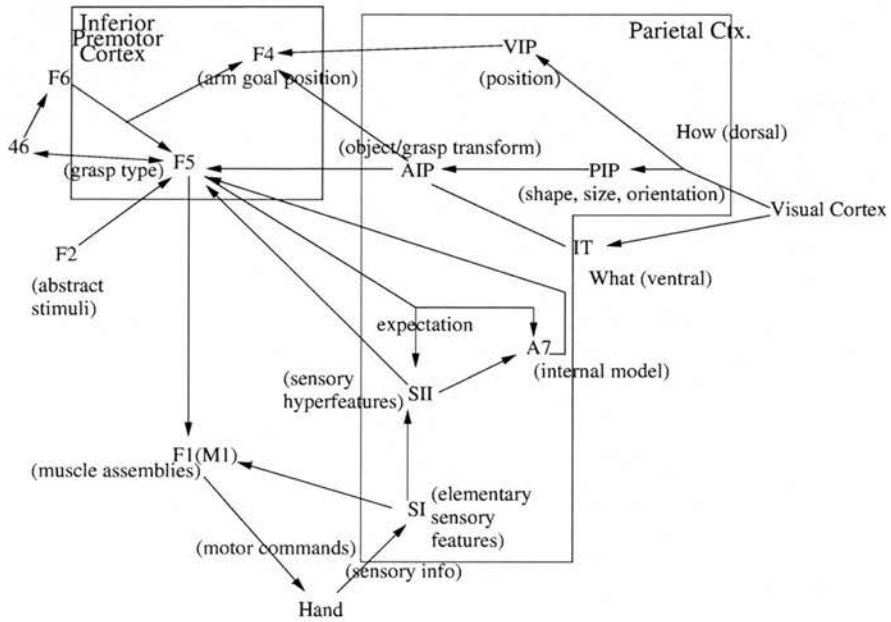


Figure 2.4: Fagg-Arbib-Rizzolatti-Sakata Model. Abbreviations stand for: AIP (Anterior Intraparietal Area), VIP (Ventral Intraparietal Area), IT (Infero-Temporal Area), PIP (Posterior Intraparietal Area). The remaining anatomical medical notation for different parts of the brain.

The FARS model is not directly concerned with behaviour selection or reinforcement learning. However, it includes an intrinsic component of interaction with the environment. Unlike this, I contend that the execution of behaviour triggers a feedback signal from the environment, which relates the three processes of interest for this thesis: affordance learning, behaviour selection and learning.

The FARS model has been used as a model for learning by imitation Rizzolatti et al. (2000); Guazzelli et al. (1998); Oztop and Arbib (2002). The process starts by areas V2 and V3 (premotor cortex) by providing visual input to the AIP (Anterior Intraparietal Area). This is used to make a coarse selection of the possible affordances of the object. Area F5 refines this selection by applying constraints depending on the task according to the signals from area F6, on the working memory (from area 46) and on the stimuli (from area F2). This relationship between F5 and AIP fits with the hypothesis of the AIP acting as an active memory of the one

selected affordance and updating this memory to correspond to the sort of grasp executed. In a complementary fashion, area F5 is held responsible for selecting a single grasp after integrating the task's constraints with the set of grasps afforded by the object nearby. The FARS model (Fagg and Arbib, 1998) provides a reasonable explanation of the way in which F5 may accept signals from areas F6 (pre-SMA supplementary motor area), 46 (dorsolateral prefrontal cortex) and F2 (dorsal premotor cortex) to respond to task constraints, working memory and instruction stimuli, respectively, and how these may be sensitive to the object recognition process in area IT (Infero-Temporal Cortex).

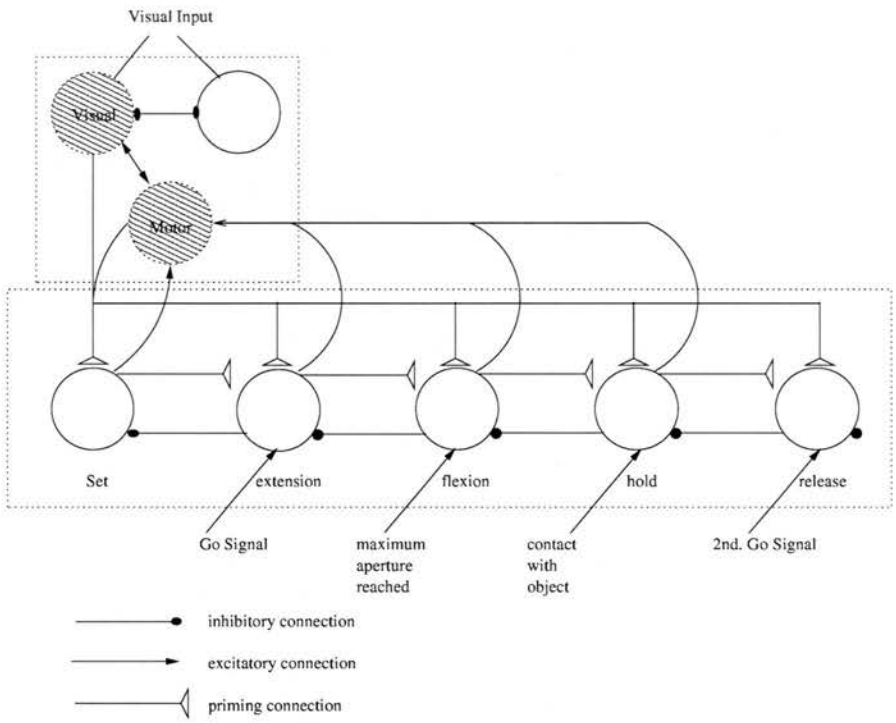


Figure 2.5: Interaction between AIP (Anterior Intraparietal Area) and F5 area neuron cell populations (FARS Model).

Related to this are also Gibson's arguments on perception. According to Gibson, the perception of object's affordances works on the basis of relating cues to behaviours, e.g., for the case of grasping: *To be graspable, an object must have opposite surfaces separated by a distance less than the span of the hand* (Gibson, 1986). This hypothesis, for the case of the grasping affordance, has been hypothesised in terms of Arbib's theory of *virtual fingers* and of *opposition space*. According to this, a grasp is defined via two "opposition axes", the *opposition axis in the hand joining the virtual finger regions to be opposed to each other*, and the *opposition axis in the object*.

This has been implemented in Fagg and Arbib (1998) by their theory of *virtual fingers* and *opposition space*. In these terms, a grasp can be defined via two "opposition axes", the *opposition axis in the hand joining the virtual finger regions to be opposed to each other*,

and the *opposition axis in the object*. If combining Gibson's and Arbib's perspectives, visual perception seems to provide the necessary *cues* to the AIP, LIP (Lateral intra-parietal area), VIP (ventral intra-parietal area), areas 7a and 7b, and area SII (second somatosensory area). Each of these areas has a *menu* of the possible motor primitives for the animal (monkey) for every action: saccadic eye movements (LIP), ocular fixation (7a), reaching (7b), and grasping (AIP). Everyone of these areas are within the brain's IPL (Inferior Parietal Lobule), receiving projections from the occipito-temporal areas and from the visual field periphery of V3 and V2. For the case of *grasping*, an appropriate opposition axis in the object can be determined after an affordance is selected, and hence the appropriate grasping behaviour can be selected and executed via mutual interaction between the F5 and AIP area, cf. figure 2.5.

In this light, I suggest that one can view the FARS' model from a more general perspective. An animal has to perform a task to satisfy each internal need by interacting with those elements in its niche offering the right affordances, e.g., a source of water when thirsty. Furthermore, this occurs in a chain of processes, starting when perceiving the environment and ending by the selection of behaviours/actions. On the one hand, sensory cortices perceive and represent sensory information, which is transferred to the basal ganglia in order to select a behaviour (Redgrave et al., 1999). According to the reward obtained via interaction with the environment behavioural patterns may be modified.

The FARS model is a very interesting model from a neurological perspective, since it assembles together the necessary processes and related parts of the brain for the grasping affordance. However, some further abstraction is required if this is to be applied to a mobile robot, since other actions than grasping will be necessary. From this perspective, I propose to abstract from this model in order to learn object affordances other than grasping. In this context, I have considered the process of *learning to perceive* object affordances. The process can be described as follows:

1. The perception of an object.
2. The relationship between the perceived cues and the possible functions of the object is estimated.
3. A behaviour or action is selected according to the repertoire attached to the object.
4. The behaviour is executed.
5. The environment returns some feedback during and after the execution of a behaviour.
6. The agent learns that the perceived cue may indicate the potentiality of executing that behaviour.

I propose to apply this method to learn object affordances in a mobile robotics' framework to improve adaptation. This is further explained in the next section.

2.8 Adaptive Perception: Learning Affordances

Gibson's classification of affordances extends to any possible action, either related to a physical movement or to a mental or abstract execution of behaviour. Furthermore, mobile robots may have to perform other actions than grasping when performing their tasks. For example, the affordance for walking is possible whenever there is a surface affording a strong enough support to sustain our weight, a picture affords to be watched, a chair affords to sit. Thus, I propose to extend the same principles introduced in the last section to other actions.

Affordance learning is intimately related to the *knowledge grounding* problem, i.e., learning that the perception of certain sensory patterns and fluctuations of the internal physiology are related to the execution of a certain behaviour. In other words, learning the potentiality of executing a particular behaviour when certain stimuli are perceived is *knowing* that this may fulfil some internal goals. Based on this, I argue that learning affordances is a different formulation of the same knowledge grounding problem, for which there is an extensive literature (Steels, 1994; Harnad, 1990). However, this has so far not been related to learning and adaptation in robotics.

The models reviewed in the first sections use external stimuli in their procedures to select actions. However, they restrict the use of external stimuli to their combination with internal drives to select behaviours, disregarding the possibility of self-learning their *potentiality*. This is exactly the extension I am proposing in order to improve a robot's adaptation to its environment. I argue that this ability is fundamental if the environment experiences a change affecting the agent's internal needs. The procedure I have designed to this end is inspired after the principles of the FARS model at a phenomenological level, and based on the assumption that this can be learnt by interaction with the environment only. The framework and the method I have used to model this are introduced in the next section.

2.9 Physiology, Ecology and Behaviour

Probably the most famous example of an artificial ecological agent is Pfeifer's Fungus Eater (Pfeifer, 1994). This is mainly based on the consideration that every creature is the result of a process of continuous interaction with the environment. Its emerging behaviour is a dynamic process, a balance between the mutual effects of the agent and the environment. This is the framework within which I propose to address the integration of affordance learning and behavioural patterns.

The set of goals of an animal defines the focus of interest in the environment and therefore, its drives for action. Complementarily, the environment constrains the manner in which the interaction takes place via physical laws. The results of these constraints are the **behaviours** of every species. For example, in the case of the simple bacterial organism *Dictyostelium*

discoideum, behaviours have adapted to respond to light gradients (phototaxis), temperature gradients (thermotaxis), pH differences (acidotaxis), and wind current (rheotaxis) in a response to the environment addressed to satisfy its nutritional and reproductive needs (Marée et al., 1999). The essential goal is to find the right conditions to reproduce; being the difference between the ideal and sensed conditions the leading force to this end. This is an example of behavioural adaptation resulting from the animal-environment mutuality.

In this light, *the role of perception consists of providing the right information to arbitrate among the behaviour repertoire* in order to satisfy the agent's goals. I can simplify this process of design by allowing the process of interaction with the environment to restrict these relationships in an analogous fashion to natural organisms. Furthermore, as suggested in the previous paragraph, the agent's perception does not need to be global, it can be restricted to the relevant elements for the fulfilment of the task. In a practical manner, this can be determined by the agent's internal goals/needs. For example, in the case of the aforementioned bacteria (*Dictyostelium discoideum*), its perception to perform phototaxis is limited to light intensity (the only salient element in the environment). Therefore, a framework consisting of a set of goals together with the interactions and the structure of the niche would be sufficient to frame a biologically resemblant agent. Evolution and adaptation will provide the boundaries for the system to exist and to survive. I propose to approach the design of the agent and its environment in an analogous fashion for the agent to be capable of adapting to its scenario; a framework to learn the affordances of an agent in its scenario.

This level of knowledge from the environment is necessary if the agent has to evolve in an environment. However, some mechanism to govern the interaction is also necessary. Otherwise, only a reactive behaviour would be possible. To this end, I argue that this perception framework needs to be embedded in a larger context to govern the interaction with the environment, a framework combining these external stimuli with the internal wills of the agent to bias the agent's behaviours in an adaptive manner: the actor-critic.

In an analogous manner to animals, these principles suggest using the structure of an animat to integrate these principles of interaction with the environment. To this end, I have based an artificial assessment system on the notion of stability proposed by Ashby (1965). The effect of a behaviour reflects on the body of the agent through the compensation of its needs. Hence, an interaction is considered to be *successful* if it contributes to compensate a need. Conversely, if it does not compensate a need it is considered as *failed*. The affordance learning criterion has to also match the sensory patterns of that behaviour when its execution is *successful*, and to unrelate them otherwise.

As a whole, behaviour selection in robotics was formally posed by Tyrrell (1993) and has been further addressed by Humphrys (1997), Spier and McFarland (1996), Avila-García and Cañamero (2002) from an ethological viewpoint. However, a neuro-ethological perspective has

only been introduced by Redgrave et al. (1999) and Guazzelli et al. (1998), whose complete model for action learning consists of a combination of DRAMA (Billard and Hayes, 1998), the MSN and the FARS model (Oztop and Arbib, 2002; Fagg and Arbib, 1998). Our view on behaviour selection draws on learning applied to behaviour selection in an alternative fashion to that proposed by Redgrave, and applied to perception in a Gibsonian manner.

Summary

The overall goal of this thesis is to build an agent capable of learning object affordances and of integrating these in an actor-critic reinforcement learning architecture. To achieve this aim, perception and behaviour selection are presented in this thesis as two dynamic processes of interaction with the environment. This animat consists of a set of processes: perceiving the right information from the environment, combining this with the expression of internal needs and executing the appropriate behaviour to satisfy the internal needs. Furthermore, this is assessed by a criterion inspired after Ashby's notion of viability (Ashby, 1965) and after the ecological principle. This relates the internal physiology to the environment and provides a guideline to implement the model described in previous sections. Next we review a set of contributions in order to justify the set of situations where this framework will be tested.

On the one hand, classical ethological models are based on the assumption that the selection of one behaviour over another works by comparison of different drives, these being an expression of internal and external stimuli. Each of them has an associated set of responses, which are activated whenever its associated drive exhibits the highest level. On the other, engineering criteria were the inspiration of robotic models, e.g. Maes (1991). However, Tyrrell's thesis pointed out the limitation of the aforementioned principles and put forward a novel inspiration for their future design. Tyrrell analysed several models, both inspired in ethology and robotics, concluding that Maes's architecture performed worse than classical ethological ones. Ever since, there have been several attempts to address the topic with mixed inspiration: Gadanho (2002), Blumberg (1997), Cañamero (1997). Some of these models were focusing on animation or basic emotions, however their most important contribution has been the identification of the relevant cognitive issues for behaviour selection. Two of these still require further study: the methods to *combine external and internal stimuli* and the understanding of methods to combine *appetitive and consummatory behaviours*. I argue that finding a solution to these problems requires novel principles of design inspired in neuroscience and ethology. Furthermore, to address these topics it is necessary to have an appropriate framework, namely a picture where the effect of behaviour selection and dynamic perception can be quantified.

Therefore, this thesis is about **ecological adaptation** inspired in principles of neuroscience and ethology. The topics addressed are:

The **first topic** addressed is the *learning of object affordances*. Their relationship to behaviour selection is dual: firstly, affordances (the agent's unit of ecological perception) and decision making are intrinsically related via the feedback from the environment, also related to **reward**. Reward is based on the effect provoked by the execution of a behaviour on the bodily physiology. Secondly, learning affordances also means to ground the knowledge from the environment with regard to the agent's internal goals. This is necessary to make efficient decisions and to be able to adapt if the affordances of the surrounding objects change.

The **second topic** addressed in the thesis is the combination of external stimuli (affordances) with the internal drives for *learning to select behaviours*. This goal is dual: on the one hand, to demonstrate that the learning of affordances has been correctly assessed, on the other to assess the aforementioned combination of stimuli to bias a heterogeneous behaviour repertoire.

physiological and behavioural, being inter-dependent. It is from this perspective that

The principles used for learning are inspired in biology and ecology, and affect both perception and behaviour selection. The novelty introduced, beyond the architecture presented in the next sections, is that this process is now **self-regulated**. Perception and behavioural patterns will vary if there are changes in the physiology or in the environment inviting to do so. For example, if the food becomes scarce, the behaviour eating may have to be selected more often to compensate this scarcity.

In this model, behavioural patterns are governed by the actor-critic, which has been hypothesised to be assessed by reward in biological systems. This, together with the Temporal Difference (TD) algorithm, address learning in an incremental manner, not needing an *a-priori* model of the environment, coherently with the ecological, developmental adaptation to the environment.

A series of **hypotheses** have been formulated to build this learning model:

- Affordance learning is driven and modulated by reward.
- The common currency for behaviour selection is reward.
- Behaviour selection is performed via comparison of motivations related to behaviours, since these are related to the predicted reward obtained via their execution.
- Learning to make decisions consists of two sub-processes:
  - Calculating the motivations related to each behaviour, predicting the expected reward (due to its execution) on the basis of external and internal stimuli.
  - Having an appropriate criterion to decide for one behaviour over another on the basis of a set of activation values.

The testing of these hypotheses is conducted under the following **assumptions**:

- The architecture for learning and behaviour selection responds to ecological principles.
- External stimuli can be modelled as affordances, i.e., reward-driven (functional) relationships between cues in the environment and the behaviours within the agent's repertoire.

The expected **results** should demonstrate that:

- Affordances can be integrated into behaviour selection.
- An actor-critic and TD provide a framework for naturally integrating perception and behaviour selection in a biologically resemblant manner.
- The principle for learning affordances is based on strengthening the relationships between cues in the environment and the motivation related to each behaviour. This relationship is the affordance of that cue with respect to that behaviour.
- The assessment criterion is based on correcting the predictions of reward for every cue with regard to each behaviour, via interaction with the environment.
- Learning to perceive from the environment and to select behaviours are part of the same problem.
- Appetitive behaviours are competing for execution in equal conditions to consummatory behaviours.

Finally, I have also argued for a single interaction dynamics, integrating physiology, the learning of functionalities and of policies. The laws of interaction with the environment will bias adaptation towards looser or stricter policies to select behaviours on demand of the environmental conditions in a way that will try to cover the need of internal, physiological stability, mandatory for the survival of the agent.

Chapter 3

A Model of Ecological Learning for Perception and Behaviour Selection

This thesis focuses on ecological learning as a process underlying perception and biologically inspired behaviour selection. The set of hypotheses and assumptions pertaining to them are addressed and disseminated throughout the different chapters composing this thesis. Nonetheless, it has been considered appropriate to collect them in this chapter as an introduction to these. Beyond introducing a loose view of model to be used in the experiments, this chapter also aims at providing an overview of the ensemble of assumptions and reasonings that relate each element of the model and that ground their interactions. The first section addresses ecological perception and the related elements of the model. This is followed by a section focusing on the parts of the model involved in behaviour selection and learning and the principles underlying their functioning. Finally, a conclusion relates elements of both previous sections and introduces the next chapters.

3.1 Grounding Affordances in the Physiology

Ecological Perception, Learning in terms of Artificial Physiology The *animat approach* (Meyer, 1997, 1995) views animats as creatures capable of "actively seeking the information they require and of selecting those behaviours that allow them to profit from their interactions with the environment." (Meyer, 1997). Different robotic architectures (Avila-García and Cañamero, 2002; Gadanho, 2002; Velásquez, 1998; Cañamero, 1997; Spier and McFarland, 1996; Blumberg, 1994) focus upon building behaviour selection architectures to make decisions according to the agent's internal state. In most architectures, although the agents can set their own goals, the relationship between the objects in the environment (or their perception) and the internal goals is usually hard-wired. However, Wilson suggested that "At each point (in building the animat) we will be careful to include *full connection with a sensory environment*,

together with maximum use of perception, categorisation, and *adaptation*" (Wilson, 1991). This is consonant with the direct extraction of information from the sensory flow proposed by Gibson (1966). In this respect, one of Gibson's main contribution to the understanding of perception is the concept of *affordance*. Affordance is the unit of perception in functional terms defined as the causal relationship between a set of cues and the potentiality for action that their perception offers to a particular agent. This is further explained in chapter 4.

According to this, it can be interpreted that this is equivalent to learning to perceive the world in an active manner. Additionally, that *learning object affordances* is one of the possible ways towards "the maximisation of the use of perception" suggested by Wilson (1991). Furthermore, a natural consequence of this ability is the conclusion that an agent capable of learning affordances would also be able to adapt to scenarios where functional relationships between objects and behaviours can be established.

However, this view is still incomplete if it remains ungrounded in the agent's internal physiological dynamics. In fact, in order to establish functional relationships between the perception of a cue and the elicitation of a behaviour defining an affordance, the effect of the execution of a behaviour on the agent's internal physiology has to be measurable. This is the principle proposed to learn the affordances of a scenario with regard to the physiology of a certain agent (affordances equally depend on the scenario as they do on the agent).

In this light, an affordance can be seen as the causal relationship between the agent's sensory flow at the moment of the interaction and the internal effect of the execution of its related behaviour. This is further addressed in chapter 4. The next section introduces the elements of design of a motivation-driven architecture as the context within which to test the aforementioned hypothesis.

Modelling Artificial Physiology and Affordance Learning The model which we present to learn affordances consists of three main elements: an *Internal Physiology module*, a *Self-Organising Feature Map (SOFM)* and a *Learning Module*.

The *agent's internal physiology*, in analogy to the motivation-driven architecture proposed by Cañamero (1997), consists of two main elements. Firstly a set of *homeostatic variables*, abstractions of the agent's internal resources. Secondly, a set of *drives* (cf. centre-left figure 3.1), which signal the level of deficit or excess of their related variables. For example, nutrition, stamina and boredom could be the homeostatic variables, and hunger, tiredness and restlessness their related drives.

The *Self-Organising Feature Map* consists of a Grow When Required (GWR) network, —cf. top-left in figure 3.1— (Marsland et al., 2002), which groups together similar sensory patterns as nodes of a topological network. The metric of similarity is the Euclidean distance between sensory patterns. The choice of the GWR network over other Self-Organising Fea-

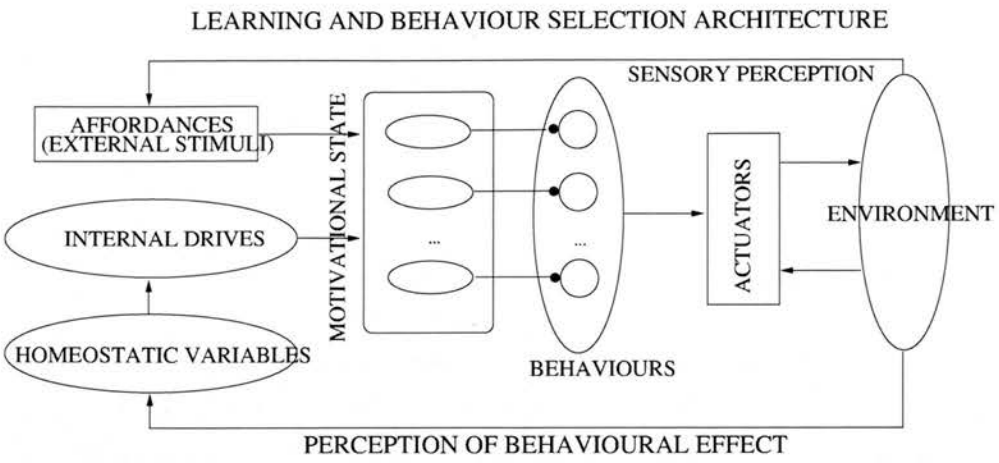


Figure 3.1: Complete Model Schema: Physiology (homeostatic variables and drives modules), SOFM (affordances, external stimuli module). Both contribute to the motivational state and to determine the intensity of the behaviours, in competition for the agent's actuators. The learning process is not explicitly indicated, but affects the connections relating the motivational state to the behaviours.

ture Maps (e.g., Kohonen networks (Kohonen, 1982)) is motivated by the ability of the GWR network to grow as much as required by the complexity of the sensory space (according to a pre-set level of accuracy). If the environment is static, each node in the network will represent a set of similar sensory patterns when its final structure has been reached. Similarly, if the environment is dynamic, the network will try to dynamically follow up its changes. If a sensory pattern is not encountered anymore, its related node will fade away and eventually disappear. Conversely, if a new sensory pattern arises, the network will create a new node to represent the novel element in the agent's sensory space.

The *Learning Module* consists of a set of synapses relating each of the nodes in the SOFM to every behaviour of the agent's repertoire. The strength value of each synapse (between 0 and 1) indicates the strength of the affordance for a given sensory pattern. The strength of the synaptic weights is established via Hebbian reinforcement (Hebb, 1949). This strengthens or weakens the weights of the synapses on the basis of the internal effect provoked by the execution of a behaviour. Behaviour execution alters the internal physiological dynamics (e.g., it increases the level of glucose in blood). This effect is signalled via a *hormonal release*. If it is beneficial, it strengthens the functional synapse that connects the behaviour just executed to the node which was active at the moment of execution (not the topological synapses of the GWR). This operation is repeated until the weights of these synapses are stable. These weights define the behaviour-affordance of that object, e.g., its edibility, or maybe how restful it is.

To summarise, this section briefly introduces a set of modules to endow an agent with the ability to learn affordances for a given set of objects in the scenario. Importantly, the *learning*

process is based on the principle of ecology, which relates the agent's interaction with elements in its scenario to the effect these may have in its internal physiology. It is argued that this is the fundamental element that allows the agent to be able to define affordances, therefore to adapt its perception to a given scenario. This argument is explained further in chapter 4.

3.2 Actor-Critic Module

Reward-Based Developmental Learning: Adapting Behavioural Patterns In addition to the ecological framework for perception, it has been previously argued that integration between different adaptive processes is a crucial issue for developing adaptive elements and systems exhibiting a higher biological resemblance and adaptive power. Towards this aim, the second issue addressed throughout the thesis is the study of reward-based developmental processes that would allow the agent to modify its behavioural patterns according to the dynamics in its scenario.

Furthermore, it is argued that adaptive learning implies modifying behavioural patterns in a manner regulated by the internal physiology, which is commanded by the reward obtained via external interaction. Behavioural and neuroscience studies support this hypothesis (Houk et al., 1995; Schultz et al., 1993) and suggest the actor-critic as a possible solution used by vertebrates to modify their behavioural patterns. This framework, together with the Temporal Difference (TD) update procedure (Sutton and Barto, 1998), addresses learning in an incremental manner, thus not requiring an *a-priori* model of the environment; this is consistent with the philosophy of ecological, developmental adaptation to the environment.

On the basis of these principles, a series of **hypotheses** have been formulated. These have been initially introduced at the end of the literature review; however, it has been considered appropriate to re-formulate them and to list them in this first experimental chapter. These will be addressed in the subsequent chapters:

- Affordance learning is driven and modulated by reward.
- The "common currency" (Redgrave et al., 1999) for behaviour selection is also reward.
- Behaviour selection is performed via comparison of the activation of the different motivations. Each motivation represents the state of activation of a related behaviour, calculated on the basis of the reward the agent expects to obtain via its execution.
- Learning to make decisions consists of two sub-processes:
  - Learning a set of activation functions (motivations), one for each behaviour, that predict expected reward (when the behaviour is executed) on the basis of external and internal stimuli.

- Having an appropriate criterion to decide for one behaviour over another on the basis of a set of activation values.

The testing of these hypotheses is conducted via modelling a model on the basis of the aforementioned principles under the following **assumptions**:

- The architecture for learning and behaviour selection responds to ecological principles.
- External stimuli can be modelled as affordances, i.e., reward-driven (functional) relationships between cues in the environment and the behaviours within the repertoire of the agent.

The concrete hypotheses we intend to test are:

- Affordances can be integrated in an ecological framework for behaviour selection.
- An actor-critic and TD learning provide a framework for naturally integrating the aforementioned principles of perception and action selection in a natural manner.
- The principle used for learning affordances is based on strengthening weighted relationships between cues in the environment and the activation of each behaviour within the repertoire of the agent. Each weight is the affordance of that cue with respect to that behaviour (see earlier comment on this).
- The assessment criterion is based on corrections of the predictions of reward for each cue with regard to each behaviour, via interaction with the environment.
- Learning to perceive the environment and to select behaviours are part of the same problem.

Finally, the need to encompass the dynamics of each element in the system has also been argued. Physiology, learning of functionalities and of policies are part of a single regulatory process orchestrated by the interaction with the environment, which will bias adaptation towards looser or stricter policies to select behaviours on demand of the environmental circumstances, and bounded by the need of internal, physiological stability, mandatory for the survival of the agent.

Model for Learning and Behaviour Selection The model for learning and behaviour selection has been developed as an extension of the ethological-based architectures proposed by (Avila-García and Cañamero, 2002; McFarland and Spier, 1997). However, unlike previous architectures solely devoted to behaviour selection, this one incorporates the possibility of modifying patterns of selection by exploiting the aforementioned principle of ecology. In fact, the effect of interacting with the environment on the internal physiological dynamics can also

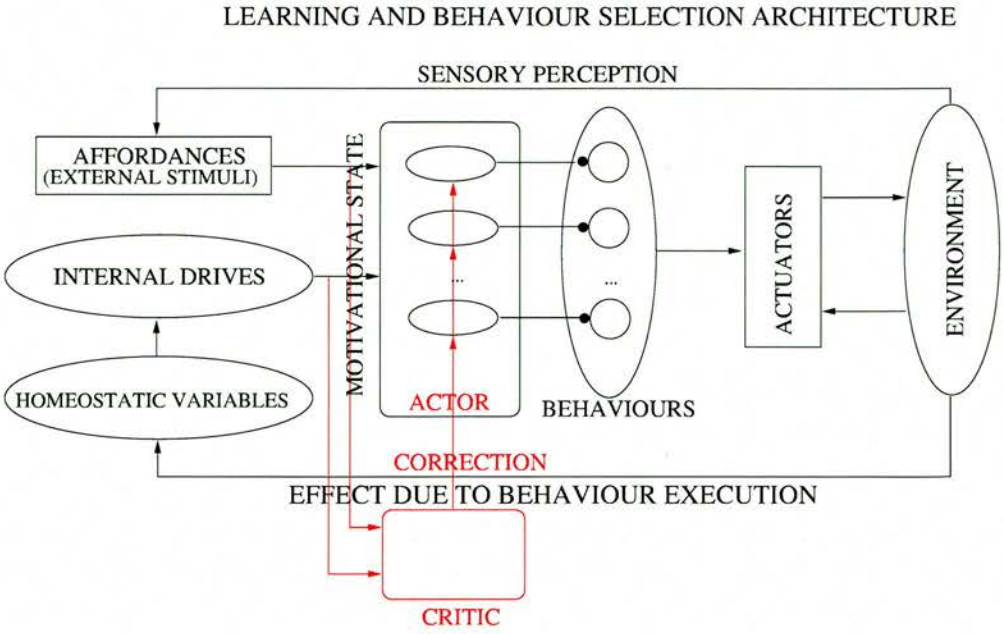


Figure 3.2: Initial Agent Model. It consists of an internal physiology, comprising the homeostatic variables and internal drives, a module to learn affordances, a motivational state, a behaviour repertoire, the agent's actuators and the environment.

be viewed in the context of reward. The elements of the architecture introduced next address the process of learning to select behaviours to maximise the prediction of reward.

The elements for learning to select behaviours are two: the actor and the critic, in addition to the modules introduced in section 3.1. As the names indicate, the actor and the critic are two different (though interacting) elements. The input to both elements is the state (motivational state), which consists of the currently perceived external stimuli and the state of the drives. As a novelty for this architecture, and unlike for others, there is no pre-defined formula to *calculate the motivational state* for every internal state. Instead, the actor-critic RL algorithm is integrated to learn to calculate these motivational states on the basis of the reward obtained by interacting with the environment.

The *actor* (cf. centre figure 3.2) decides the behaviour to execute next in such a manner that reward is maximised. This principle of learning is applied to combine external and internal stimuli to appropriately select behaviours. The selection is a competition for dis-inhibiting the behaviour which is presumably the most rewarding.

The *critic* (cf. centre-bottom figure 3.2) assesses decisions and calculates the *error in the prediction of reward* (effective reinforcement) once the decision has been made. At that moment the prediction of reward can be compared with the real reward obtained in order to correct the prediction.

This architecture for learning has a dual goal. To show that learning to perceive and to select behaviours are a sub-set of a more general reward-based learning system, in the overall

context of adaptation. Furthermore, it also aims at using an actor-critic in a simulated environment providing elements to disambiguate different theories on the main role of the basal ganglia by suggesting that both learning and behaviour selection can partly share the same neural substrate.

Summary

This chapter has introduced the different interacting modules in the model (cf. figure 3.2) and an overview of their inter-dependence and interaction. The parts of the model are:

- a module of internal resources, the level of which is signalled by the homeostatic variables. These variables express their deficit or excess via a set of related drives. This is introduced in the next chapter; section 4.1.
- the module of affordances, which provides the agent with information about the possibility of executing certain actions with the elements situated in its close scenario. This is also introduced in the next chapter; section 4.2.
- the module for behaviour selection and learning, composed of the actor critic. This is described in chapter 5.
- the behaviour repertoire of the agent.

Each aspect of the model and an extended explanation of their principles are introduced in detail in the following chapters.

Chapter 4

Ecological Perception

Biology has become increasingly popular throughout the last decade as a source of inspiration on adaptation-related ideas for mobile robotics (Avila-García and Cañamero, 2002; Gadanho, 2002; Velásquez, 1998; Cañamero, 1997; Spier and McFarland, 1996; Blumberg, 1994; Braitenberg, 1984). I argue that this line of research is justified by the fact that the best example of an adaptive system is an animal in its niche. This view as inspiration for engineering purposes has been however admonished (Hallam, 1998), since a direct application of biological principles does not guarantee an acceptable solution from an engineering perspective. Natural adaptation occurs if there are competitors for the same resources or if changes in the environment oblige; these can be viewed as the natural optimisation criteria. Unlike these, optimisation in engineering is driven by criteria imposed by a designer, which depending on the application do not need to maintain any resemblance to biology. Nevertheless, if the goals pursued by the natural and the artificial beings are alike, it seems appropriate to learn from the examples delivered by nature and from the principles of interaction that lead to themselves as autonomous individuals. This is the intention throughout this thesis.

The relationship between the individual and its environment is summarised by the ecological principle (Pfeifer, 1994) which considers an agent and its environment as two entities integrated in a single dynamics governing their mutual interaction. Interaction can be viewed from the perspective of an agent as a loop, initiated by the perception of the environment, followed by some internal interaction with its physiological values which lead to a behavioural response to modify the environment and the agent's own internal physiology. Therefore, from an ecological perspective it seems reasonable to focus on the three processes that seem to initially drive adaptation: perception, interaction with the environment and internal physiological dynamics, since adaptiveness is an emergent phenomenon of the dynamics established by these processes. In this context, this chapter addresses the study of the first of these processes: the underlying mechanisms of ecological perception. To assemble agents capable of dynamically adapting to a variety of environments firstly depends on this.

Perception in the animal realm is designed as a means for survival. It is a process endowing animals with *the appropriate information on their environment* to intelligently *manage their internal resources* (McFarland, 1993). Most animals perform sequences of actions to satisfy their needs by interacting with the appropriate elements in their environment. However, animals do not interact with everything; they filter much of the sensory information while still remaining capable of navigating, evolving and adapting to their environment. This raises questions related to the ingraining of perception in this process. How do animals perceive their world? How do they combine perception with their internal wills to drive themselves to satisfy their internal needs? How do they increase their knowledge of the environment? This chapter intends to suggest answers to these questions and to do so within the constraints imposed by the ecological principle.

It is important to stress that this is a study on *developmental learning* in the context of adaptation. Traditionally, adaptive processes have been classified as developmental and genetic (Fisher, 1930). The former modify the perception of the animal by interacting with its environment in the course of the life of the individual (McFarland, 1993). The latter prune the individuals unfit to deal with their environment and extend over generations. However, this classic separation seems to fade in the light of newer hypotheses arguing the influence of behaviour on the genome (Ackley and Littman, 1991); therefore suggesting a relationship between both processes as complementary mechanisms for adaptation. However, this is beyond the scope of this research. For practical reasons, this thesis solely focuses on *developmental learning*.

For this study on ecological perception, I have considered observation of perception and interaction processes in the animal world. However, this does not suffice to explain how animals learn to perceive their environment. Beyond perception and behaviour selection, learning also requires assessment (Sutton and Barto, 1998); hence being able to distinguish the beneficial from the harmful. Only those animals encouraging actions whose effect is beneficial and discouraging those whose effect is harmful survive. This internal assessment criterion has been commonly referred to as the *sense of valence* (Ackley and Littman, 1991). It works at a physiological level, providing an overall assessment about the changes in the physiological state due to the execution of a behaviour. For example, consuming good food when it is hungry reflects on a feeling of satisfaction because it increases the level of glucose in blood, whereas putting a hand on fire provokes a feeling of discomfort, related to the pain and harm resulting from this action[1]. The sense of valence increased the likelihood of survival of some

[1]Why are animals endowed with this sense? The most likely reason is that those lacking this appreciation do not survive in hazardous environments. Hence, they were pruned from the genetic tree (Darwin, 1866). In this respect, it is useful to view the interaction between the animal and its environment in terms of a systems interaction, whose level of organisation is granular and hierarchical, i.e., an autopoietic matrix (Maturana and Varela, 1980). In this context, the *sense of valence* can be interpreted as a basic element of cognition, a principle necessary to allow the autopoietic entity (the living being) to be, and to be in position of contributing to perpetuate the species.

species; however, the underlying physiological phenomena that gave origin to this are still uncertain. However, the need for homeostasis (Cannon, 1929), re-formulated as Ashby's notion of viability (Ashby, 1965) does provide a practical guideline for assessing the effect of the behavioural responses on the agent. Only animals whose physiological needs are kept within certain boundaries (within the viability zone) survive. In engineering terms, this principle has been the inspiration for Avila-García and Cañamero (2002) who defined a set of viability indicators to assess the performance of different behaviour selection architectures. However, I argue that this is furthermore the assessment principle guiding the learning processes, which I propose to apply to ecological perception. The reward of the artificial agent in this thesis will have an embedded assessment system dictating that actions whose physiological effect leads towards a physiological state of less deficit are rewarding. Those with the converse effect are considered as punishing.

Furthermore, it is also important to introduce the idea that this learning process has relevant implications from a conceptual perspective. Among the most mentioned processes in developmental psychology is the process of *knowledge grounding* (Noble, 1998; Vogt, 2001). I argue that the *ecological principle* can be used to establish a semantics of the objects in the environment. If the meaning of an object is its functional definition, learning the physiological effect of a behaviour related to some sensory cues is learning a possible semantics of these sets of cues at an individual level.

In this light, *this chapter addresses the study of the underpinnings of the learning processes that enable an agent to interact with objects for a given set of environments.* The level of description of these processes lies at the level of functional systems. In loose terms, I intend to endow an agent with the capability of *knowing what to do with the things it has around given the aforementioned constraints.* This is addressed by assigning significance to the cues perceived with respect to each behaviour and to each goal of the agent, hence, by endowing the agent with the capability of learning its own affordances[2]. This is reviewed in chapter 2.

The next section introduces the principles that enable us to model an agent's internal physiology, which is necessary to allow the agent to drive its actions and to model the effects of a behaviour. This is followed by a section on clustering methods, which introduces different algorithms that can be used to identify and group similar cues in the environment. The learning mechanisms are then described, followed by a section introducing some experiments carried out to test the learning algorithms. A summary of the results of these experiments precedes the discussion and the conclusion.

[2]This term was defined by Gibson (1966) as the *"functionality offered by a cue or set of cues in the environment to a particular agent"*. We suggest that *learning the affordances of an object is equivalent to grounding the knowledge of the functionality of objects with respect to the agent's internal goals if the assessment criterion respects the aforementioned assumption.*

4.1 Artificial Physiology

This section introduces the elements composing the agent's internal physiology. These provide the necessary context to understand the integration of the feedback from the environment into the agent's internal dynamics and therefore the role that perception plays in the adaptation process.

Homeostasis (Cannon, 1929) is the self-regulation of the body's *internal milieu*[3] (Bernard, 1878). From a synthetic perspective, Cañamero (1997) proposed to model internal resources as a set of *Homeostatic Variables*, abstractions of internal resources, which may express their status of deficit or excess via *drives* (Hull, 1943) or *motivations*. These need to be satisfied either by internal compensation or by external action (behaviour execution).

For the latter case, the environment provides the necessary feedback for this satisfaction or compensation and restricts via physical laws the manner in which interaction is possible.

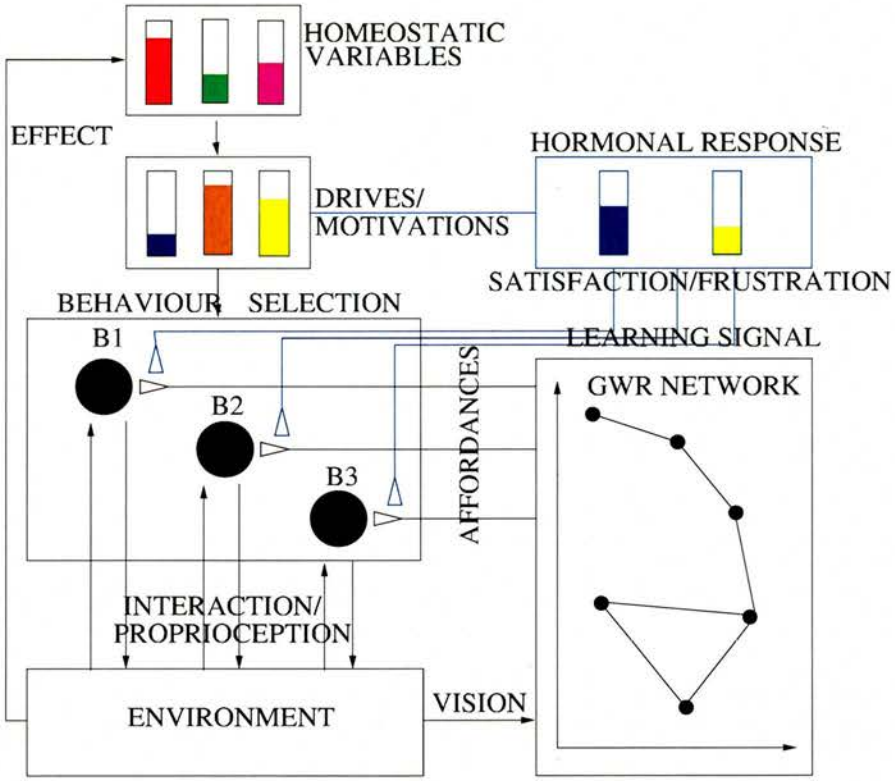


Figure 4.1: Affordance Learning and Behaviour Selection Model. It consists of a set of homeostatic variables, a set of drives and motivations, a behaviour repertoire, a hormonal set and an affordance learning module.

The overall goal for any living organism is survival, which directly relates to Ashby's definition of physiological balance (Ashby, 1965). The physiological deficits of an agent bias the execution of behaviours, the completion of which depends on meeting and recognising the

[3] Agent's internal physiology.

appropriate resources. Given these processes, I propose to analogously view an artificial agent as a set of interacting elements as shown in figure 4.1: a set of homeostatic variables (top left of figure 4.1), a set of drives (underneath the homeostatic variables in figure 4.1), a hormone satisfaction and frustration (on the right hand side of the drives in figure 4.1) and a behaviour repertoire (B1, B2 and B3 in figure 4.1) and an affordance learning module (based on a GWR network and a set of synapses relating perception to every behaviour) – bottom right in figure 4.1 –. The structure of every module and the dynamics of the overall agent are introduced next:

The controlled homeostatic variables (top-left corner in fig. 4.1) are abstractions representing the agent's internal resources, e.g., stamina, nutrition, restlessness. Their values must be kept within the *viability zone* (Ashby, 1965) for the agent to remain alive; if their values overflow/underflow the upper/lower boundaries of the variable, the robot 'dies'. Each homeostatic variable can have a status of 'normality', excess or deficit. Their behaviour is governed by the equation

$$\dot{V}_i = -C_i + \sum_j \alpha_{ik}\delta(t - t_j). \tag{4.1}$$

In equation 4.1, the homeostatic variable V_i decays over time with $\tau_i = -1/C_i$ if there is no feedback from the environment. α_{ik} is the amount of effect relating behaviour k to variable i. The summation over j indicates the contributions from the environment to the compensation of variable i. These contributions are made each time a behaviour is executed (time t_j). In the general case, every homeostatic variable can be compensated by more than one behaviour (and so I would also have to sum over all behaviours k). Nevertheless, it has been initially considered to reduce the complexity of the system by relating the agent's internal physiological variables to the behaviours one to one. Therefore, for this case every behaviour compensates a single homeostatic variable. For example, if some food is ingested, the value of its related variable (nutrition) is increased by α_{ik} (see table 4.1 for a set of typical values). $\delta(t)$ is the Dirac-delta function.

The second element of the physiological model introduced in this thesis are the *drives* (top-left corner in figure 4.1). If the role of the homeostatic variables is to represent the agent's internal resources, the drives express the deficits of these resources, quantified as a function of the differential between the optimal value and the instantaneous value for every related homeostatic variable. Equation 4.2 is the generic equation for a drive:

$$D_i(t) = \sum_j a_{ji}(V_{op_j} - V_j(t)) + \sum_k b_{ki}\dot{V}_k. \tag{4.2}$$

D_i is drive i and the V_i and V_k are the values of the related homeostatic variables. V_{op_j} is the optimal value of the j^{th} homeostatic variable, a_{ji} and b_{ki} are the coefficients relating the variable and its derivative, respectively, to the drive. The initial simplification for the modelling of the drives has been to restrict the number of homeostatic variables that influence every

drive, furthermore that the drives are implicitly independent of the derivative of the homeostatic variables. Therefore, the summation for the first expression will solely contain a single term, and all values for b_{ki} will be zero. Therefore, when a homeostatic variable diverges from its optimal point, an appropriate mechanism of compensation is triggered. In this case, the mechanism of compensation is the selection and execution of a behaviour.

The arbitration mechanism for behaviour selection follows a winner-take-all policy, using the drive that exhibits the highest urgency (the one with the highest level) to choose the behaviour to execute next. In the first experiment set, it has been organised in such a manner that a single behaviour can satisfy every drive. The amount of compensation potentially provided by behaviour k to variable i at the time t_j is expressed via the term $\alpha_{ik}\delta(t-t_j)$ in equation 4.1. Every behaviour execution will only succeed (α_{ik}[4] will be larger than 0) if the object nearby affords that behaviour to be executed. For example, hunger (controlling nutrition) needs an edible object, fatigue (controlling stamina) a resting object, curiosity (controlling restlessness) any object to interact with. At every time step, each drive is assigned an intensity proportional to the magnitude of the error of its controlled variable, as shown by equation 4.2.

The behaviours (cf. centre-left in figure 4.1) are in this case grasp, shelter, rest; unless it is specified otherwise. Behaviours are coarse grained and include a subset of actions. For example, for the behaviour to have a compensatory effect (increment or decrement on the internal homeostatic variables), its execution must happen in an agent-object framework exhibiting that affordance, e.g., if the agent is a human and the object a door handle, the behaviour to open the door may be executed. The increments to each variable as a result of the successful execution (some drive is diminished as a result of this execution) of a behaviour are introduced in table 4.1. Any successful interaction affects the agent's internal deficits[5], and the way in which this is done is shown in figure 4.2 (Spier and McFarland, 1996), where $d(t-1)$ and $d(t)$ are two vectors representing the values of the drives (every dimension relates to a drive) before and after the execution of the behaviour. The vector linking them is the effect vector, the components of which are α_{ik} for every dimension k.

The model also includes two *hormones*, satisfaction and frustration, which are internally triggered when there is a sudden variation of the internal homeostatic variables and exhibit an exponential decay after this moment. Their role is to strengthen or weaken the neural structure that learns object affordances, as described in section 4.3.

The rest of the architecture introduced in figure 4.1 is the sensory pattern clustering module (labelled as GWR module) on the right hand side of the figure. Its role in the perception process together with the elements relating perception to the agent's internal physiology will

[4]The α_{ik} have been fixed *a priori* of every trial to the values introduced in table 4.1.

[5]For the case represented in figure 4.1, there are only two deficits, one on each axis, and a viability zone defined by some boundaries that the agent has to maintain itself within to survive. The execution of a behaviour can have two different effects, either it does not modify the agent's internal state (failed execution) or it diminishes the level of one or more deficits (successful execution).

var_i/$behaviour_k$	*Grasp*	*Shelter*	*Touch*
nutrition	0.3	0.0	0.0
stamina	0.0	0.2	0.0
restlessness	0.0	0.0	0.1

Table 4.1: Effects of each behaviour on the homeostatic variables (α_{ik})

be explained in the next section.

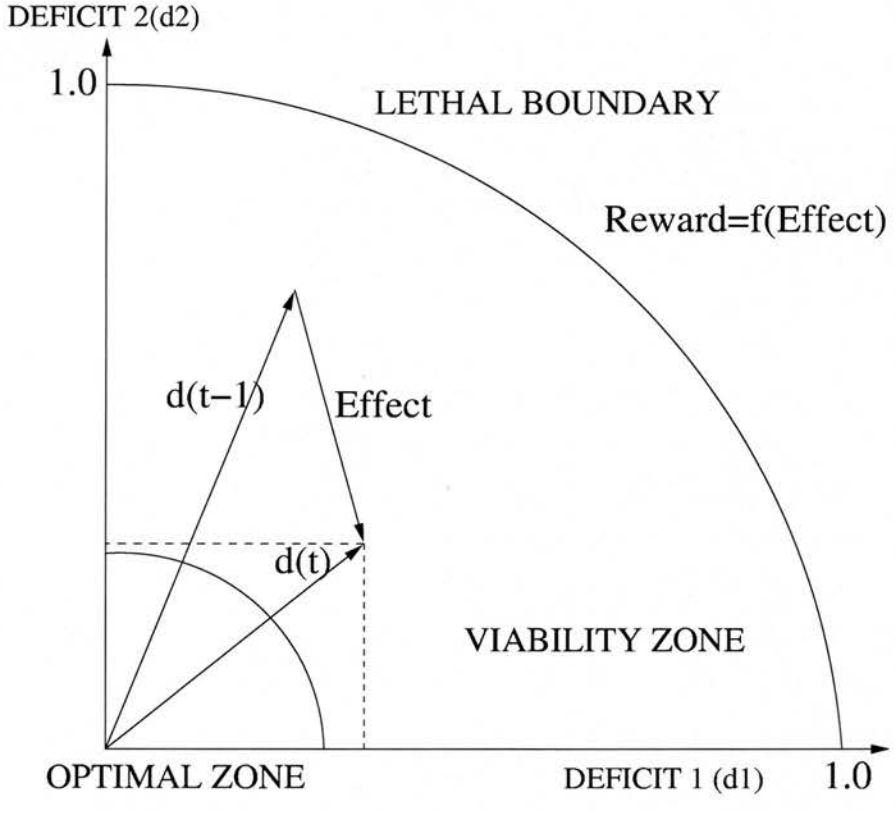


Figure 4.2: Depiction of Reward in Physiological Terms in a 2-D physiological space. $d(t-1)$ and $d(t)$ stand for the physiological state before and after the execution of a behaviour. The vector *effect* stands for the amount of physiological compensation due to its execution. Optimal zone is the area where the deficits are minimal. Viability zone stands for the area of tolerable deficits. This is closed by the lethal boundary.

4.2 Ecological Perception

It was previously stated that the goal of this study is to model the mechanisms for an agent to learn to perceive the environment in an ecological manner. To this end, the previous section has introduced the set of elements that explain the dynamics of the agent's internal physiology.

In a complementary fashion, this section introduces the assessment system and the procedure to group together sensory patterns from the environment as an intermediate step for learning affordances.

4.2.1 Feedback from the Environment

The model introduced (cf. figure 4.1) aims at learning affordances by relating the perception of sensory cues (*invariants* in the agent's sensory flow) to fluctuations of the internal physiological dynamics. The *affordances* are modelled as synaptic relationships between the representation of clustered sensory cues and every behaviour within the agent's repertoire. The learning process then consists of strengthening those synapses whose related behaviour execution leads towards a higher physiological stability and to weaken those evoking the converse effect. This is indirectly driven by the hormonal response; satisfaction reinforces synapses, frustration has the converse effect. Therefore, learning is based on the *interaction between the agent's physiological dynamics and the environment.* This is further explained in section 4.3.

This is analogous to biological behaviour, since the interaction with the environment provokes a proprioceptive (response in the environment that can be perceived by the agent's sensory apparatus) and a hormonal response detectable at different time-scales (Kravitz, 1988). By learning from this hormonal feedback, the agent will be able to *anticipate* the outcome of an interaction; and will therefore be able to decide whether it is worth carrying out that interaction with that object or if it is preferable to search for an alternative. Initial experiments with a simulated Khepera robot demonstrated that this results in a better adaptation to the environment in terms of life span (Cos-Aguilera et al., 2003).

The model introduced to learn object affordances comprises several parts: a clustering module to extract patterns of the sensory flow, an architecture for behaviour selection to choose the behaviour to execute next and a learning module. The principles underlying the behaviour of Self-Organising Feature Maps as clustering devices are introduced next.

4.2.2 Growing Networks

Animal perception systems can distinguish stimuli with a level of precision determined by the tasks they execute for survival (Pfeifer, 1994). Furthermore, animals lack of a sense of aesthetics and are indifferent about the features of the stimulus as long as they offer the right functionality. For example, horses do not seem to care about the shape or size of the apples, as long as they are edible.

Therefore, stimuli can be classified according to their physical properties and to their relationship to the agent. In other words, perception seems to classify objects according to their physical similarity and to their functionality. Furthermore, this occurs as an *adaptive* process, since animals respond to *novelty* by integrating new combinations of sensory patterns in their

neural representation, the affordances of which will have to be inferred. Thus, an adaptive artificial implementation of this will have to be extendable.

Two Self-Organising Feature Maps (SOFM) exhibit the required properties: Growing Neural Gas (GNG) network (Fritzke, 1994) and Grow When Required (GWR) network (Marsland et al., 2002). Both networks relate their connectivity to the geometrical similarity of the sensory signals used to train them, measured by a given metric. Furthermore, unlike Kohonen networks (Kohonen, 1982), for these cases the number of nodes grows proportionally to the complexity of the sensory space and does not need to be fixed a-priori. New nodes are inserted when the mismatch between the sensory pattern and its closest node rises over a certain threshold.

These neural algorithms dynamically adapt to the level of entropy of the sensory signals and they do so in an incremental manner — commonly used synapses are strengthened, conversely for those seldom triggered, which tend to fade and to disappear in the long run. The main difference between both is that the GNG network adds a node periodically, while the GWR network does so only when the topological representation is considered to be not accurate enough.

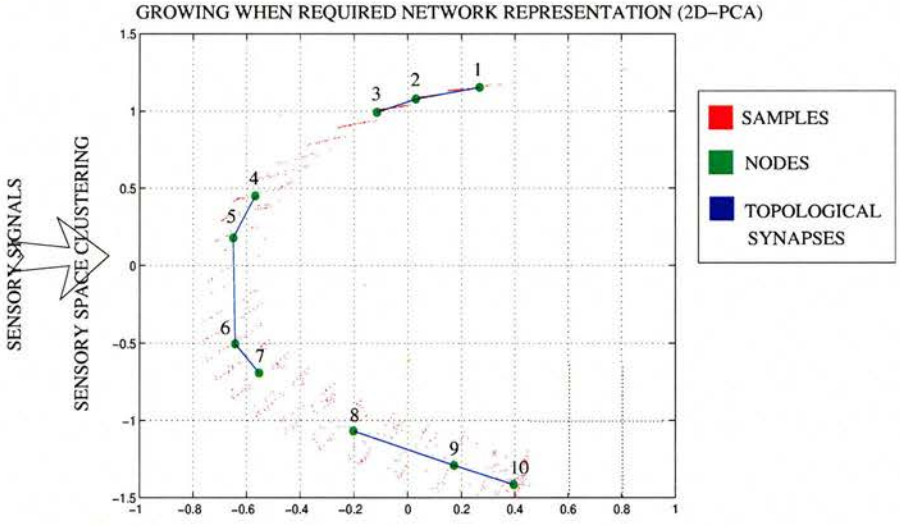


Figure 4.3: 2-Dimensional Principal Components of a generic Topological Map (SOFM). The GWR network clusters the sample space represented by the red dots. The network is represented by the green nodes, connected by the blue synapses.

The SOFM adapts the topology of its nodal representation to the patterns of the sensory signals. The representation therefore organises in sets of nodes, which can be identified and numbered, grouping sets of objects exhibiting similar perception in a single representational unit —cf. figure 4.3 green dots. Furthermore, Gibson's ecological approach argues that animal perception works in a functional manner by using the regularities of the optical flow to elucidate the potential for action in that situation (its affordances). Thus I hope that the SOFM will organise to capture these regularities. The schema introduced in figure 4.1 shows an overview

of the first implementation of this ecological view.

The next sections address the growing of two different types of SOFMs, a GNG network and a GWR network. The former is trained with object features, which the agent perceives from the objects in its surroundings, the latter are trained with raw sensory signals. These two different approaches reflect two complementary views on how perception is built. On the one hand, it has been suggested that sets of neurons in the vertebrates' neural substrate (V1 to V5 areas) are sensitive to objects' features (size, shape, orientation) (Tao et al., 2004). On the other, Gibson argued that raw sensory signals are directly related to the agent's behaviours. Based on these two perspectives, the two aforementioned experimental approaches have been proposed.

4.2.3 Feature Based Perception

The first set of experiments, introduced in section 4.4.1, addresses the use of object features, e.g., size and shape, to build a topological map. A Growing Neural Gas Network was used to this goal (Fritzke, 1994).

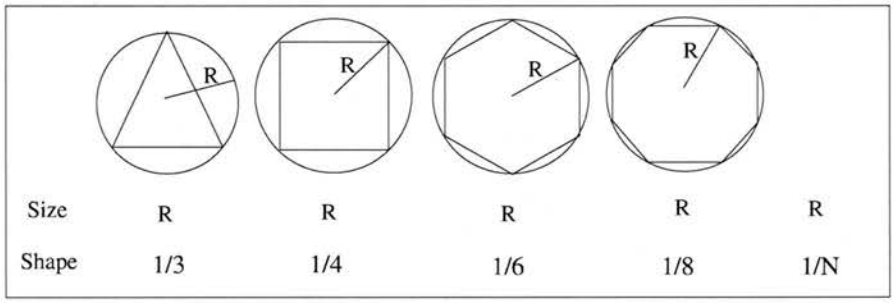


Figure 4.4: Features Definition for Common Objects: shape and size are the features considered.

For the first two sets of experiments, the GNG network has two pre-processed signals as inputs; the size and shape of the object closest to the agent. The size of the object is the radius of the object's base and the shape is computed as the number of sides of the object's polygonal base, cf. figure 4.4. The GNG (Growing Neural Gas) network is initialised with two nodes only; located at positions ω_k and ω_l of the sensory space, chosen at random. At each presentation of a sensory signal ξ to the network, the algorithm behaves as follows:

- Locate the two nodes i and j closest to the samples by calculating the Euclidean distance from the sample ξ to each node ω_i

$$\Delta e_i = \| \omega_i - \xi \|^2 \ \forall i \in \{1..N\}, \tag{4.3}$$

where N is the number of nodes.

- The winner (i.e., closest) node i and the nodes immediately adjacent in the topology are shifted towards the location of the sample by an amount proportional to the mismatch between them. The winner node is shifted by

$$\Delta_i = \varepsilon_b(\xi - \omega_i) \tag{4.4}$$

and the nodes immediately adjacent by

$$\Delta_k = \varepsilon_k(\xi - \omega_k), \forall k \in \{1..L\}, \tag{4.5}$$

where L is the number of adjacent nodes.

- The age of the synapse between nodes i and j is re-set to zero. If it does not exist, it is newly created.

- Synapses the age of which is larger than a_{max} are deleted together with nodes with no synapses.

- Every λ samples a new node is inserted equidistantly between the two units, q and f, with the largest accumulated error. If there is a synapse between these two nodes, it is deleted and two new ones connecting them to the newly inserted node are created. Furthermore, the error of units q and f is decreased by a factor of α and the new node is initialised with the average error of nodes q and f.

- All error variables are decreased multiplying by parameter d.

These steps are repeated *ad infinitum* or until a criterion of convergence is reached. In this study, a final number of nodes has been pre-set. Four examples of GNG network are shown in figure 4.8, corresponding to the worlds represented in figure 4.7. The GNG network adapts its topology by shifting and adding nodes and synapses to the locations where input signals are most frequent. If there is an area of the sensory space where the samples rarely appear, nodes or synapses located within it will age and be deleted in the long run. This is the algorithm used to cluster sets of object features.

In a complementary fashion, the second possibility consisted of directly clustering raw sensory signals. This procedure is performed in section 4.4.3 in four different scenarios. This issue is further extended in the next subsection.

4.2.4 Raw Sensory Data Perception

The first set of experiments clustered combinations of pre-processed features, which could be obtained by using feature detectors (Kohonen et al., 1997). It is argued that this intermediate step facilitates the clustering, since processed features are easier to classify; however, though

correct from the perspective of learning affordances, ecology argues in favour of a principle of economy, which can be better supported if meaningful knowledge is directly extracted from the sensory flow. For example, while an object is approached, the sensory flow contains much information, from which only a few cues may be significant to perform that behaviour. From this perspective, I introduce the following set of experiments (see section 4.4.3), where perception is based on raw object images. The sensory inputs in this case are snapshots taken always at a fixed distance from the object. We now describe how I have used Marsland's Grow When Required (GWR) network. The vector associated with each node is a 64×64 image of 8-bit pixels.

- Analogously to a GNG network, the first and the second closest nodes to the sample, nodes i and j, respectively, are selected —the metric is the Euclidean distance[6].
- A synapse between i and j is grown (if not yet existing).
- The activity of the best matching unit is calculated according to

$$a_i = exp^{\|\omega_i - \xi\|^2},\ i \in \{1..N\}. \tag{4.6}$$

  where ξ is a sample from the sensory input and ω_i the vector of node i. This is the quadratic power of the original activity metric proposed by Marsland. In this way, samples which are close will be considered to be even closer, and those which were separated will be considered to be further apart. If the activity is lower than the threshold a_T and the habituation threshold is lower than h_T, a new node k is inserted between the best matching node i and the sample ξ. The new node k is connected by two new synapses to nodes i and j and the synapse between i and j is deleted. Its weight vector is the average of i and χ.

- If no new node is added (the activity of the winning node a_i is smaller than a_T and/or the firing threshold h_i is larger than h_T), the winner node i and the nodes immediately adjacent in the topology are *shifted* towards the location of the sample by an amount (ε_b and ε_k) proportional to the mismatch between them. The winner node is shifted by

$$\Delta_i = \varepsilon_b \times h_b \times (\xi - \omega_i) \tag{4.7}$$

  where ω_i is the vector of node i. The nodes immediately adjacent are shifted by

$$\Delta_k = \varepsilon_k \times h_k \times (\xi - \omega_k),\ \forall\, k \in \{1..L\}, \tag{4.8}$$

  being L the number of adjacent nodes, ξ a sample from the sensory space, h_b and h_k two normalisation constants and ω_k the vector of node k.

[6]This is defined as $\Sigma_{i=0}^{N}(x_i - y_i)^2$.

- The age of all nodes connecting to the winner node are incremented by 1.
- Reduce the habituation according to

$$h_i(t) = h_0 - \frac{1}{\alpha_i}(1 - exp(-\alpha_i t/\tau_i)) \tag{4.9}$$

where α_i is a normalisation constant and τ_i the decay constant. Likewise, the counters of its neighbours k respond to

$$h_k(t) = h_0 - \frac{1}{\alpha_k}(1 - exp(-\alpha_k t/\tau_k)) \; \forall \, k \in \{1..L\}. \tag{4.10}$$

where α_k is a normalisation constant and τ_k the decay constant.

- If there are nodes with no synapses or synapses the age of which is larger than age_{max}, these are deleted.

The *parameters* of the network are as follows. Firstly, *activity threshold* a_T, secondly the *habituation threshold* h_T. The h_T is a value signalling the boundary in time allowed for a single node to place itself in its best fitting location. Beyond this boundary, it is not allowed for the node to move any further to represent the data more closely; hence a new node is needed. The third parameter is the maximal *age of synapse* age_{max}. Synapses connected to frequently used nodes are re-set to 0, conversely, a natural ageing is experienced that beyond a threshold leads to the deletion of the synapse and of the nodes (nodes with no synapses are also deleted). The final parameters are the *shifting coefficients* ε_b and ε_k; which specify the dragging speed of the nodes towards the new sample. Despite working on a statistical basis, the GWR offers the advantage of controlling *novelty* and *habituation* via the parameters presented in this paragraph.

So far the methods for fitting the agent's sensory space have been described as an introductory step to learn object affordances. The next section introduces the learning method for establishing functional relationships from sets of cues represented by a node in the SOFM to behaviours.

4.3 Learning Method

The problem posed consists of learning to relate the perceived regularities to the potentiality of performing a behaviour. To this end, I propose to grow *functional* synapses, represented by the letter χ_{ij} (cf. figure 4.5), which differ from the topological synapses (ω_{jk}) inherent to the topology of the SOFM. The functional synapses χ_{ij} indicate the potentiality of performing behaviour j when node i is active. Therefore, if the agent is able to establish these weights, it will be able to predict the outcome of performing certain behaviours when its related node is perceived. The process is illustrated in figure 4.5 and is described next:

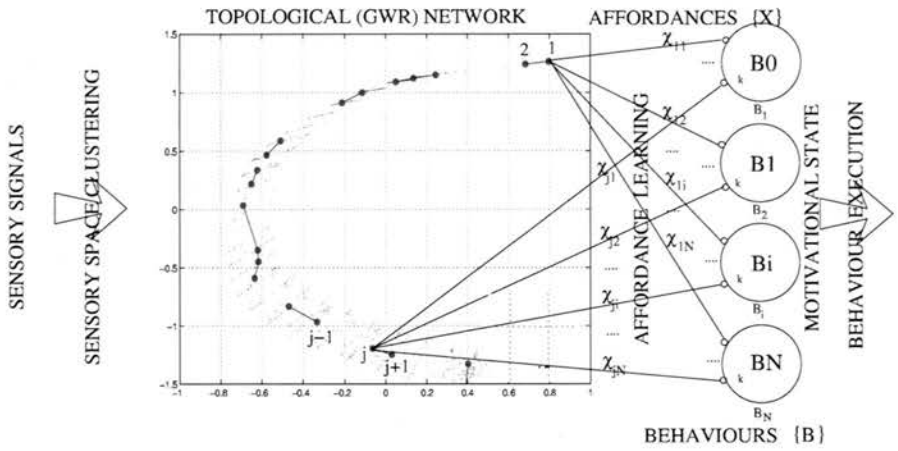


Figure 4.5: Affordance Learning Framework. χ_{ij} stands for the synaptic weight relating behaviour i to the node j of the SOFM.

- At the detection of an object, the closest node (Euclidean distance) is identified.
- If the interaction succeeds there will be a reduction of one or more of the deficits that will trigger the hormone satisfaction. Conversely, the hormone frustration will be triggered. The release of one or another hormone conditions the growing or fading, respectively, of the functional weight (likelihood) that relates the node active at that moment to the executed behaviour according to

$$\chi_{ij} \leftarrow \chi_{ij} + \alpha b_j. \tag{4.11}$$

  where χ_{ij} is the weight between node i and behaviour j, where b_j is the intensity for behaviour j (cf. right hand side in figure 4.5). It is 1.0 if j is the behaviour just executed and 0.0 otherwise. Therefore, only the weights relating the active cue i to behaviour j are updated by a value α. This is positive if the hormone satisfaction has been released and negative otherwise.
- The learning results in a set of weights relating each cue in the environment (represented with a node in the network space), to each of the behaviours. These are the affordance values of the object represented by the nodes for that particular agent.

In a step-by-step fashion, the learning process is as follows:

1. Growing the SOFM. The network is trained with a series of sensory patterns[7] to which it shall adapt its structure. The adaptation algorithm compares the pattern (red dots in figure 4.5) to the node space of the network (green dots in figure 4.5): if the Euclidean distance between the closest node and the current sensory pattern is considered to be too

[7] A pattern is a sample of the input sensory signals.

far away, a new node is inserted equidistantly and new synapse (blue lines in figure 4.5) is added. Conversely, the closest node and the nodes in its neighbourhood are dragged towards the input pattern to better represent it. Nodes seldom close to the input pattern are deleted. A fully detailed description of the growing algorithms is provided in the previous section and in Fritzke (1994); Marsland et al. (2002) for the GNG and GWR networks, respectively.

2. Identifying the Nodes. Once the network has been grown and has a stable structure, the nodes of the network are numbered. For the GNG network case, the nodes have been assembled in clusters (each set of inter-connected nodes is considered a cluster), identified by a centre (the mean of the position of each node in the cluster) in the input space, and a measure of its dispersion (the distance between the centre and the furthest node). For the GWR, single nodes are treated and numbered individually. Numbering is necessary to be able to relate the sensory cues to each of them during the growing of the functional synapses relating to the behaviours.

3. Use of the Network. Once the network has reached a stable topology, it can be used for object identification by identifying the closest node to the sensory input.

This learning process has been engineered in analogy to biology, being that the *baby* agent starts to recognise discrete objects and to attach some functional value to their perception. This argument is further extended in section 2.6.

This learning procedure is Hebbian (Hebb, 1949). However, it is guided by a reinforcement signal, see expression 4.11. Several interaction episodes happen repeatedly throughout the duration of the simulation and modify the values of the functional synapses relating the nodes in the SOFM to the agent's behaviours. The final values are associated to the probabilities of succeeding when executing a behaviour and have been normalised between -1.0 and 1.0 at the end of the learning process. These measure the matching between the node (the regularity in the environment) and the behaviour potentials in an analogous fashion to a normalised probability value between -1.0 and 1.0.

This learning mechanism has been extensively studied in a series of environments, varying the perception from features to raw sensory data, and from the GNG to the GWR networks. These experiments are introduced in the next section.

4.4 Experiments

The set of experiments introduced next has been performed with a simulated Khepera robot with a *Webots*^{TM} simulator 4.0. The overall goal of the robot in these environments is to survive by learning to interact with the objects in its surroundings. To this end, the robot is

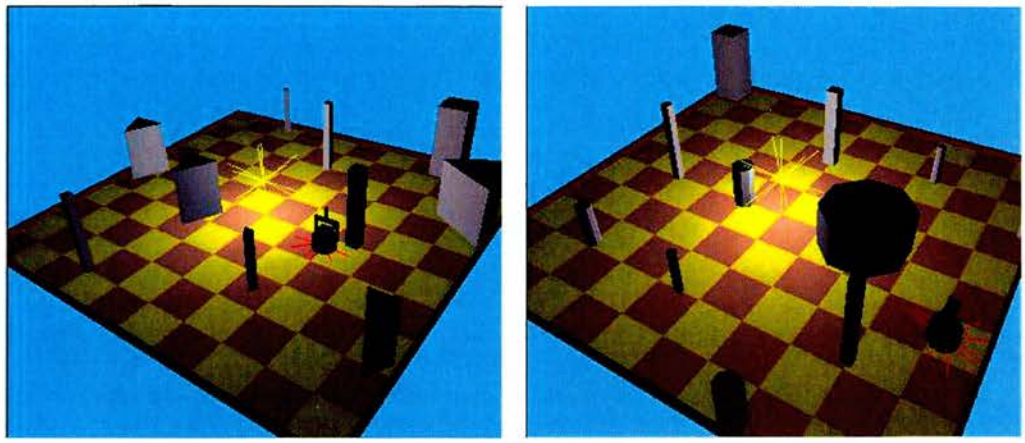


Figure 4.6: Worlds 1 and 2 (left and right, respectively). World 1 contains a series of polyhedral objects with triangular and square shapes, and world 2 a series of polyhedral objects with octagonal and square shapes. The number of objects in each scenario is the same (12 objects each).

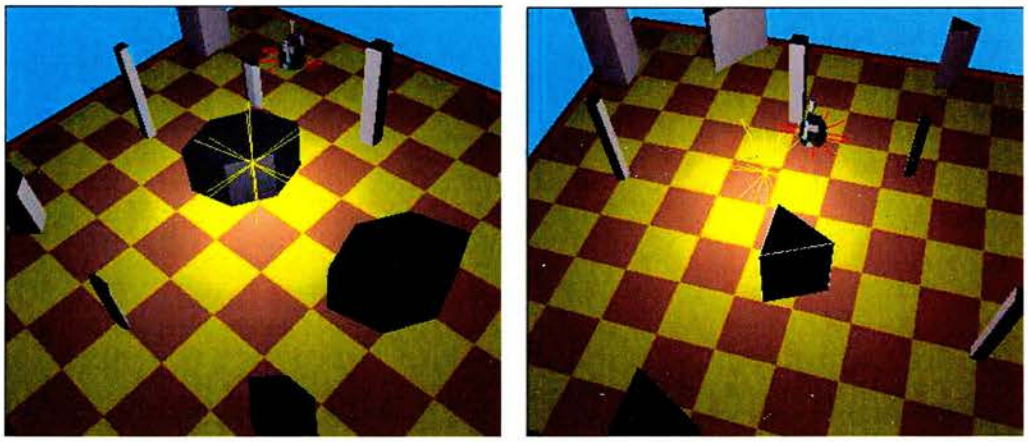


Figure 4.7: Worlds 3 and 4 (left and right, respectively). World 3 contains a set of octagonal and hexagonal shapes and world 4 a set of triangular and square shapes. The number of objects in each scenario is the same (12 objects each).

programmed to continuously wander and interact. At every encounter with an object, the robot has the opportunity to execute a behaviour, which may compensate an internal need depending on the affordances offered by the object. Once an interaction with an object has occurred, the object is abandoned to search for a new one.

4.4.1 Feature Based Perception Experiments

The goal of the first experiment set is to demonstrate that the artificial agent can learn to relate object features, represented by clusters in the GNG network, to the outcome of goal-oriented in-

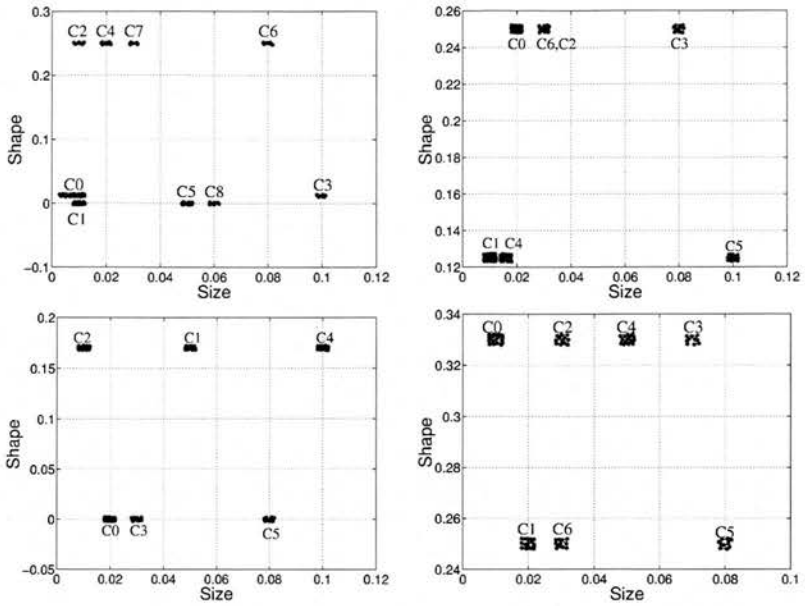


Figure 4.8: Clusters for Worlds 1, 2, 3 and 4 (top left to bottom right, respectively). These are environments containing the shapes specified in figure 4.7.

teractions. This would be analogous, loosely speaking, to learning to select appropriate objects (or the equivalent set of regularities) to successfully perform an interaction[8]. The environments shown in figure 4.7 which generated the clusters represented in figure 4.8 have been used for simulation.

The *experimental method* is described next:

1. The *GNG Network* is built in order to match different patterns of features until its structure becomes stable. The parameters used to grow the network are: $a_T = 0.5$, $\varepsilon_b = 0.5$, $\varepsilon_n = 0.006$, space dimension=2, $a_{max} = 50$, $\lambda = 100$ and D=0.995. A thorough description of the algorithm and of the aforementioned parameters can be found in section 4.2.3. Gaussian noise has been added to the input signal. The position of each of the nodes for each cluster is stored when the stable representation is reached, cf. figure 4.8.

2. The *homeostatic variables* of the agent are initialised to their optimal value. The homeostatic variables for these experiments are nutrition, stamina and restlessness. These are controlled, respectively, by the drives hunger, fatigue and curiosity. Table 4.2 shows the exponential decay parameters for each homeostatic variable, their optimal set points and their range of variation.

3. The *arbitration mechanism* selects the behaviour whose related drive exhibits the highest value at the beginning of each interaction episode. Each of these drives is satisfied via

[8]By successful I mean the interaction exerting a positive effect on the agent's internal resources.

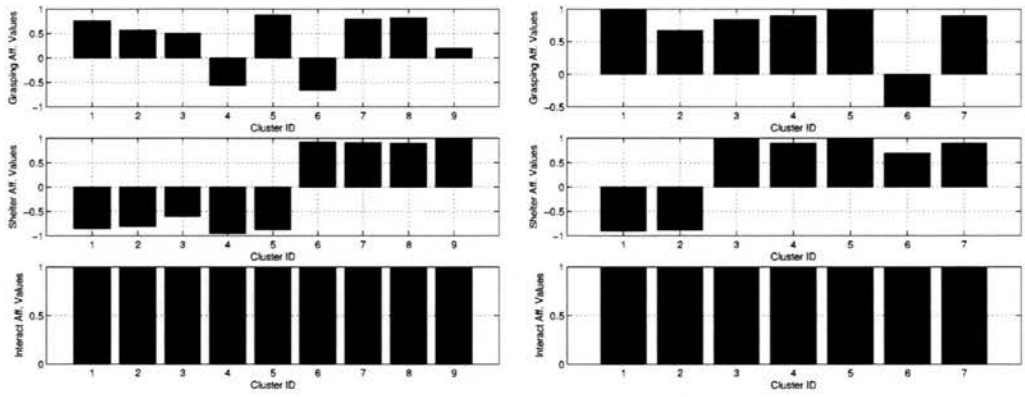


Figure 4.9: Reinforced Values for Clusters in Environment 1 and 2, left and right, respectively. Cluster ID correspond to the clusters labelled on the two top representations in figure 4.8. The values represent the affordances that relate them. The value 1.0 means that an object close to that cluster affords the behaviour with probability 1.0. Conversely, if the value is -1.0 the behaviour is not afforded.

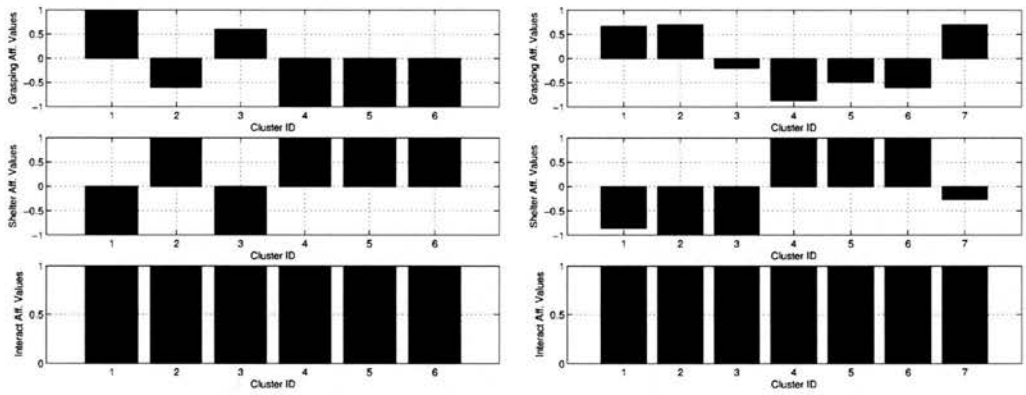


Figure 4.10: Reinforced Values for Clusters in Environment 3 and 4, left and right, respectively. Cluster ID corresponds to the clusters labelled on the two bottom representations in figure 4.8. The values represent the affordances that relate them. The value 1.0 means that an object close to that cluster affords the behaviour with probability 1.0. Conversely, if the value is -1.0 the behaviour is not afforded.

executing a behaviour, to which it is related, "nutrition" to grasp (eat), "stamina" to shelter and "restlessness" to interact, respectively.

4. Once an object is encountered, the pattern is compared to every neural node in the topological representation in order to locate the closest one.

5. A behaviour is selected at random and executed for exploration purposes. If the features of the object were appropriate, the interaction will be successful. In this case, the home-

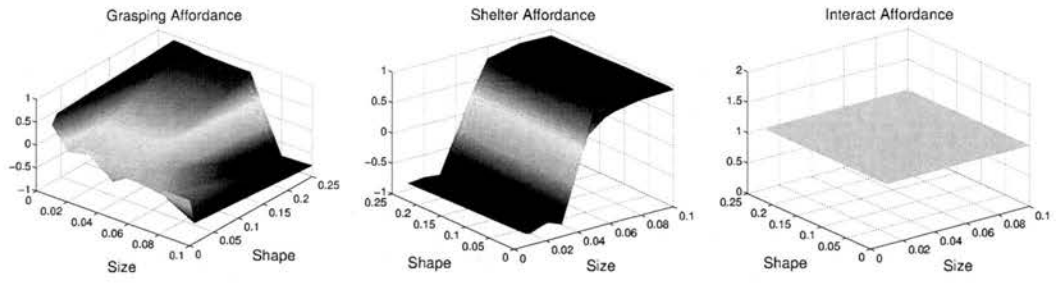


Figure 4.11: Grasping, Shelter and Interact Affordances (from left to right, respectively) for World 1. Values close to 1 mean that the behaviour is afforded, conversely, values close to -1 mean that it is not afforded. Depending on the shape and size of the object, other intermediate values have been obtained, resulting in these gradients.

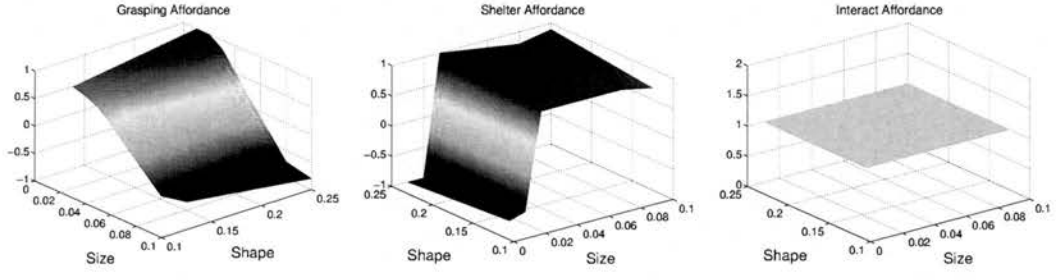


Figure 4.12: Grasping, Shelter and Interact Affordances (from left to right, respectively) for World 2. Values close to 1 mean that the behaviour is afforded, conversely, values close to -1 mean that it is not afforded. Depending on the shape and size of the object, other intermediate values have been obtained, resulting in these gradients.

	Nutrition	**Stamina**	**Restlessness**
τ	*1E-4*	*-1E-4*	*1E-4*
Opt. Value	*0.8*	*0.2*	*0.3*
Range	*[0.0,1.0]*	*[0.0,1.0]*	*[0.0,1.0]*
Drive	*Hunger*	*Tiredness*	*Restlessness*
Behaviour	*Grasp*	*Shelter*	*Interact*

Table 4.2: Homeostatic Variables: Simulation Parameters. τ is their decay constant; Opt. Value stands for their respective set points. Each variable is continuous and ranges between 0.0 and 1.0.

ostatic variable compensated due to the performance of a behaviour moves towards its optimum. If the interaction fails, it has no physiological effect[9].

[9]Unlike for the first set of experiments published in Cos-Aguilera et al. (2003), where a negative outcome implied a negative impact on the level of the homeostatic variables.

6. Depending on the success or failure of the interaction, the satisfaction or frustration hormone is triggered to strengthen or weaken the functional synapse between the active node and the behaviour just executed.

This process is repeated in simulation over 500,000 steps (time units in the simulation environment) for each scenario. 20 simulations of each time have been performed in order to obtain statistically significant results. These are explained next.

4.4.2 Results

This experiment set has been tested in the four scenarios illustrated in figure 4.6 and 4.7, each of which contains objects with different shapes and sizes. Each scenario contains two different sets of objects, each with different features. Their resulting GNG networks have been presented in figures 4.8 and their resulting histograms in figures 4.9 and 4.10. The values in the histograms represent the likelihood of successfully performing a behaviour when that node is active. Values range between -1 and 1 (1.0 means the maximum likelihood[10] and -1.0 its minimum). The columns specify the node (labelled as 1 to 7) and the rows the behaviour. For illustrative purposes, I have also drawn the interpolated values for each behaviour (and affordance likelihoods) in figures 4.11, 4.12, 4.13 and 4.14. The x and y axes represent the shape and size of the perception state, respectively, and the z (vertical) axis the resulting affordance value.[11]

The left hand side graphs in figures 4.11, 4.12, 4.13 and 4.14 show that small objects afford to be grasped more or less independently of their shape. Objects larger than 0.04 (the width of the Khepera's gripper) do not afford to be grasped. A threshold is set for every scenario. Objects whose size is larger than this will not provoke a compensation of the drive hunger when the behaviour is executed. Consistently, the resulting likelihood values in the middle graphs in figures 4.11, 4.12, 4.13 and 4.14 show that only objects larger than this threshold afford to shelter. The easiest figures to interpret are those representing the likelihood of an object to afford interaction (cf. right hand side graphs in figures 4.11, 4.12, 4.13 and 4.14). All objects offer this affordance, therefore the likelihood values equal 1.0.

These results demonstrate that this learning mechanism relates the agents' perception units (the cues) to the potentiality of executing each behaviour within the agent's repertoire. It can be said that agents could learn to "anticipate" the effect of executing a behaviour by using the functional weights χ_{ij} as predictors of the potentiality of performing a behaviour. Every execution modifies the neural synapse estimating the likelihood of successfully performing that

[10] This likelihood ranges between -1.0 and 1.0. In order to consider equal to the classical probabilistical concept of likelihood, it should be normalised between 0.0 and 1.0.

[11] The continuity of the 3D representation is only for illustrative purposes; it does not imply that the concept of affordance is continuous as well.

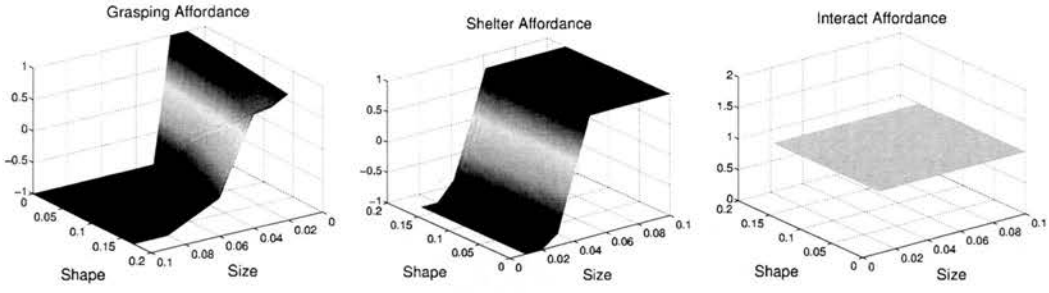


Figure 4.13: Grasping, Shelter and Interact Affordances (from left to right) for World 3. Values close to 1 mean that the behaviour is afforded, conversely, values close to -1 mean that it is not afforded. Depending on the shape and size of the object, other intermediate values have been obtained, obtaining these gradients.

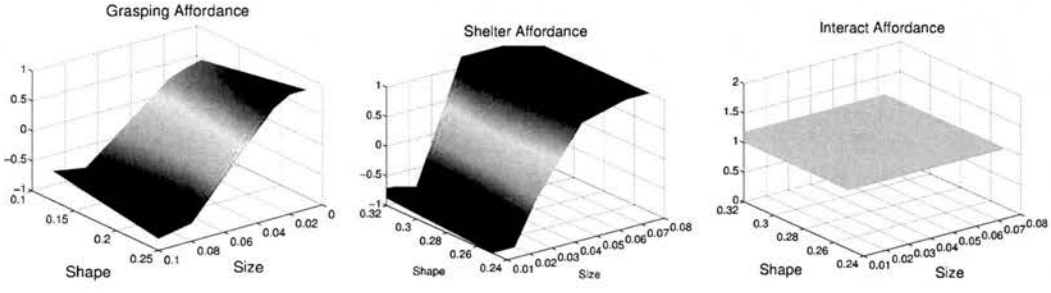


Figure 4.14: Grasping, Shelter and Interact Affordances (from left to right) for World 4. Values close to 1 mean that the behaviour is afforded, conversely, values close to -1 mean that it is not afforded. Depending on the shape and size of the object, other intermediate values have been obtained, obtaining these gradients.

behaviour. The final values of the neural synapses are the result of a statistical growing mechanism and are based on the physiological effect of performing each behaviour. These synaptic weights represent the affordance values in the framework of that agent and that particular scenario.

Beyond their predictive power, these affordances can also be viewed as a particular formulation of the grounding knowledge problem that binds together the agent and the environment in dynamical terms. This has the inherent consequence of establishing a semantic base solely valid in the context of that agent and that environment.

This first experiment set has highlighted the possibility of learning object affordances on the basis of their features (shape and size). Nevertheless, despite the theoretical soundness of this paradigm, its application necessitates the use of feature filters (Kohonen et al., 1997) designed *ad hoc* to this end. Related to this, different types of neurons in the human brain have been identified as resonators for stimuli exhibiting particular features, such as vertical or horizontal lines (Tao et al., 2004). Nevertheless, since the application is robotics, it may also be

interesting to use raw sensory flow and to let the robot define its own features and their related relevance. This idea is explored in the next section.

4.4.3 Clustering Raw Sensory Data

The possibility of clustering object features has been addressed by the previous set of experiments. This is useful from a theoretical perspective, since the principles formulated in this framework firstly demonstrated that it is possible to learn object affordances. However, directly perceiving object features is unrealistic for a robot environment unless feature filters are used (Tao et al., 2004). Instead, it is also possible to formulate a framework where a robot makes sense of its own environment by relating fluctuations of its internal physiology to other fluctuations in its sensory flow. This section explores this possibility.

Therefore, this second set of experiments tests the learning principles in an architecture to learn object affordances from raw sensory data. In order to simulate sensory flow, it would be necessary to use a temporal stream of images at varying distances which would then would have to be segmented. It is outside the scope of this thesis to do that. Instead, the sensory signals will consist of steady images taken at a fixed distance from the objects. These pieces of sensory information have been used to train the topological network.

This choice is coherent with the framework proposed (see explanation in the previous experimental setup). The two approaches, feature based and sensory input based, differ in that the nodes represent different things; for the former experiments combinations of sensory features, for this case, sets of similar sensory patterns.

Furthermore, for the former experiments, a GNG network has been used to cluster the sensory space. However, this network has been openly criticised for several reasons (Marsland et al., 2002). Firstly, this network does not directly respond to novelty. Instead, it adds a new node every λ time steps, independently of the convenience of doing so. Unlike this, the GWR network monitors the distance between the sample and its closest node and the level of habituation of this node. Two parameters are controlling this in a straightforward manner; the activity parameter a_j and the habituation threshold h_i, respectively. These parameters condition the network to insert a new node. However, unlike the GNG network, the final decision on inserting a new node depends on the distribution of the sensory signals. Secondly, it is straightforward to understand that the *accuracy* of the topology is set via balancing the habituation h_i of its nodes, controlled by its decay constant τ_i and by the maximum activity a_T (see section 4.2.4). If the dispersion of the sensory data is low, the behaviour of the GNG and GWR networks will be similar. However, the GWR network is more appropriate if the data exhibits a high or unknown level of variability, since it can dynamically increase or reduce the number of nodes on demand of the data it is trained with (Marsland et al., 2002).

The *experimental method*, analogously to the previous experiments, consists of two phases:

1. *Training the SOFM.* The robot is placed in a scenario surrounded by objects with different shapes and sizes, cf. scenario in figure 4.8. Snapshots of the objects are taken when the robot is centred and in front of an object. The data used to train the GWR are 64 horizontal element vectors of grey-level intensity, normalised between 0 and 1.

2. *Affordance Learning.* Each node in the GWR network is linked to each behaviour by a synapse initialised to a small random value.

   - A steady image taken at a constant distance from the object is used to identify it by selecting the closest node in the GWR (in terms of Euclidean distance).
   - A behaviour is selected at random and executed.
   - The effect of behaviour execution reflects internally on the level of the hormone satisfaction for the case of successful interaction (one or more deficits decrease). Conversely, the hormone frustration will increase to 1.0 (the default value for both of them is 0.0).
   - The release of the hormone satisfaction or frustration strengthens or weakens the synapse relating the behaviour just executed and the node just active according to the Hebbian algorithm, cf. equation 4.11.
   - The simulation stops when a minimum of 15 interactions per node are performed. The final values are normalised.

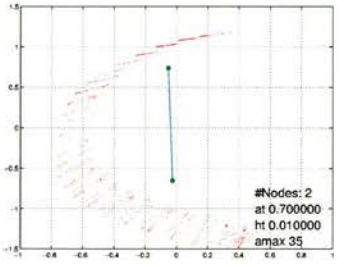
There are two different *scenarios*, a simple one containing two different objects (see section 4.4.4) and a more complex one containing a large number of objects (see section 4.4.5).

The results for each of them are described in the next sections. Two different *metrics* are proposed to assess the performance of the SOFM:

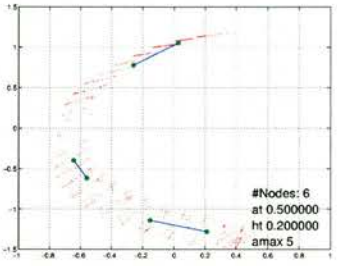
1. *SOFM Accuracy.* 10,000 image samples have been used to grow every SOFM, and 5,000 more have been used for testing purposes. The goal of this is to have a metric that allows an easy choice of parameters given a required level of accuracy. This metric consists of the *mean fitting error* specified by

$$m_{x_i} = \frac{1}{N} \sum_{i=0}^{N-1} \| 1 - x_i \| \tag{4.12}$$

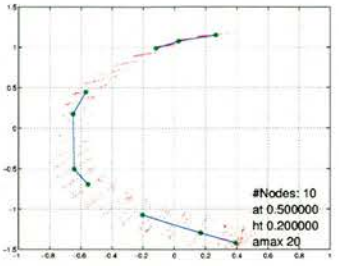
   where x_i is the weight of node i (their expected final values are +1 or -1, affords the behaviour or not, respectively) and N the number of nodes; and the related *variance fitting error*



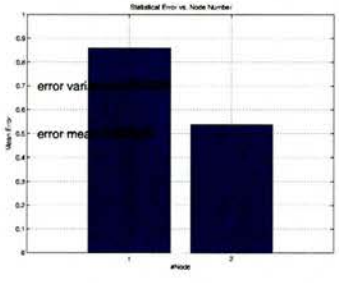
(a) 2N-GWR; $a_T = 0.7$, $h_T = 0.01$ $age_{max} = 35$



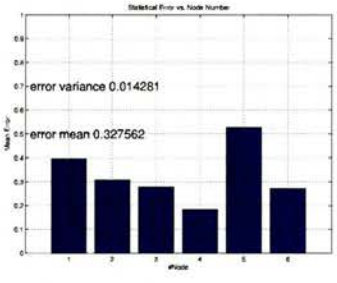
(b) 6N-GWR; $a_T = 0.5$, $h_T = 0.2$ $age_{max} = 5$



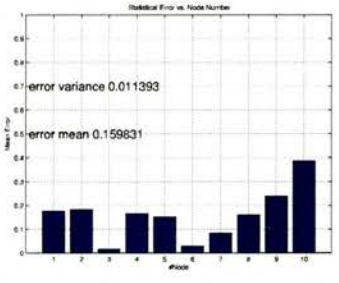
(c) 10N-GWR; $a_T = 0.5$, $h_T = 0.2$ $age_{max} = 20$



(d) 2N-GWR; Fitting Error.

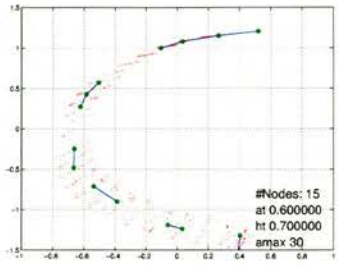


(e) 6N-GWR; Fitting Error.

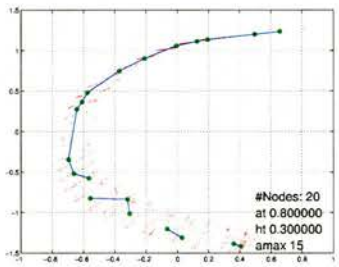


(f) 10N-GWR; Fitting Error.

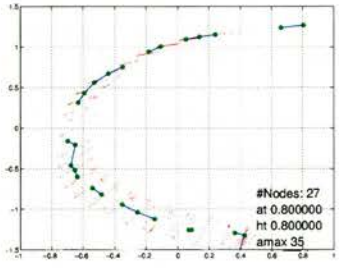
Figure 4.15: These are three GWR networks of the simple environment with 2, 6 and 10 nodes, from left to right, respectively. The red dots are the samples used for training. The resulting GWR consists of the green nodes connected by the blue synapses. The histograms underneath show the mean fitting error of the sensory space for each of them.



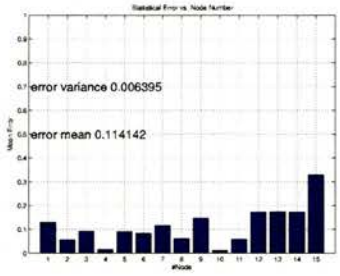
(a) 15N-GWR;$a_T = 0.6$, $h_T = 0.7$, $age_{max} = 30$



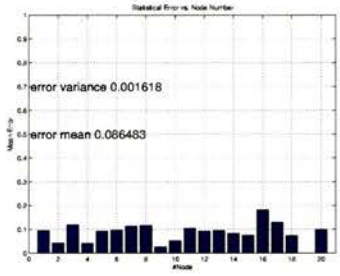
(b) 20N-GWR; $a_T = 0.8$, h_T 0.3, age_{max}=5



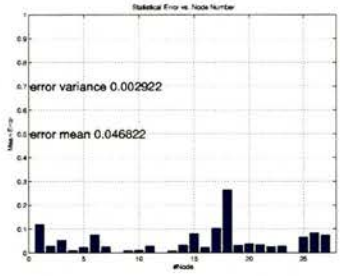
(c) 27N-GWR; $a_T = 0.8$, h_T=0.8, age_{max}=20



(d) 15N-GWR; Fitting Error.

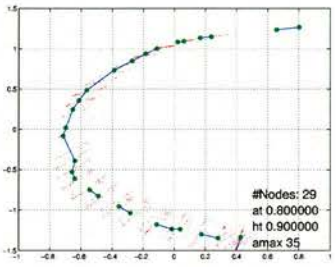


(e) 20N-GWR; Fitting Error.

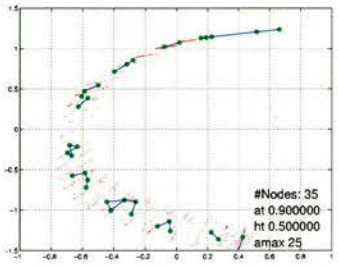


(f) 27N-GWR; Fitting Error.

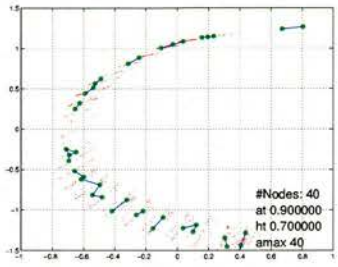
Figure 4.16: These are three GWR networks of the simple environment with 15, 20 and 27 nodes, from left to right, respectively. The red dots are the samples used for training. The resulting GWR consists of the green nodes connected by the blue synapses. The histograms underneath show the mean fitting error of the sensory space for each of them.



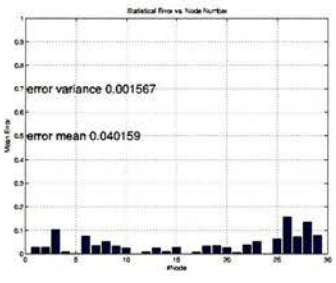
(a) 29N-GWR;$a_T = 0.8$, $h_T = 0.9$, $age_{max} = 35$



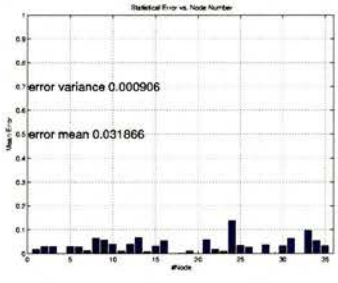
(b) 35N-GWR; $a_T = 0.9$, h_T=0.5, age_{max}=25



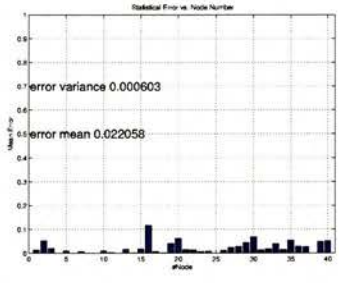
(c) 40N-GWR; a_T=0.9, h_T=0.7, age_{max}=40



(d) 29N-GWR; Fitting Error.



(e) 35N-GWR; Fitting Error.



(f) 40N-GWR; Fitting Error.

Figure 4.17: These are three GWR networks of the simple environment with 29, 35 and 40 nodes, from left to right, respectively. The red dots are the samples used for training. The resulting GWR consists of the green nodes connected by the blue synapses. The histograms underneath show the mean fitting error of the sensory space for each of them.

$$\sigma_x = m\{\sum_{i=0}^{N-1} \sigma_{x_i}\}. \tag{4.13}$$

Equations 4.12 and 4.13 refer to the calculation of the mean and variance values, respectively, of the *reliability* of the affordance values. $m\{\}$ is the mean value of the summation of error variances for every node of the GWR. Both metrics are a statistical characterisation of the affordance values obtained. The affordance values x_i are calculated as an average of successful and unsuccessful interactions with the set of objects clustered by node i in a process driven by reward. For example, x_i equals 0.7 if the interaction was successful 7 times out of 10. Ideally, the affordance values should be either 1.0 or -1.0, meaning the prediction of executing a behaviour or not with full certainty. Therefore, it is reasonable to use the metrics proposed by the equations above to assess the fitting of the GWR network. These metrics are printed on the top-left corner of each histogram (cf. figure 4.15 to 4.17, described below).

2. Since the learning mechanism relates fluctuations of internal physiological dynamics to the agent's sensory space, it has been considered appropriate to use also *viability indicators* to measure the behavioural performance of the learnt affordance values (*weight values* that relate the nodes to the behaviours). Two indicators have been chosen, namely the *physiological stability* and *overall comfort*, as defined in equations 4.14 and 4.15, respectively. These are measured throughout simulation once the affordance values are stable.

$$Physiological\ Stability = \frac{1}{N}\sum_{i=0}^{N-1} \hat{d_i}(t) \tag{4.14}$$

d_i represents one of the N deficits of the agent.

$$Overall\ Comfort = \frac{1}{N}\sum_{i=0}^{N-1} \sigma\{d_i(t)\}. \tag{4.15}$$

Where σ stands for the variance and d_i stands for deficit (or drive) i. These are similar indicators to those defined by Avila-García and Cañamero (2002) and also respond to the need of having a behavioural measurement that relates the behavioural performance to an internal indicator physiological balance. This will be further explained in chapter 5.

These two metrics respond to the two processes requiring assessment: the statistical fitting (growing) of the GWR and the performance of the learning procedure that grows the functional

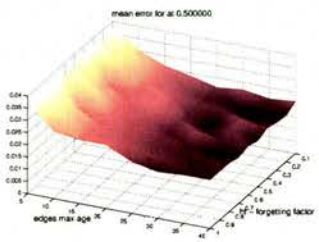
synapses between the nodes of the GWR and each behaviour within the agent's repertoire (that learns the affordance values). It is important to notice that the *computation of the functional weights* considers the real interaction between the object and the robot's sensors and effectors, the former prone to error in data collection for object identification, the latter prone to error due to inaccurate proprioception and motor-command execution.

4.4.4 Learning Affordances in a Simple Environment

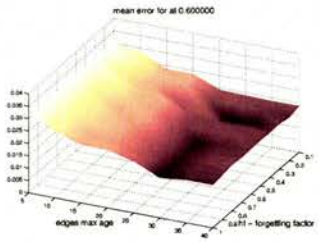
The *simple environment* contains only two types of objects, a small and a large object. This environment is used in order to assess the performance of the GWR in an easy test-bed. Is it really feasible to use a GWR to cluster objects to learn object affordances? Are the metrics proposed above significant? Is the learning principle correct? We need to address these questions before further utilising this approach.

To this end, the *first experiment set* with raw sensory data aims at elucidating the GWR parameters (activity a_T, habituation threshold h_T and maximum synaptic age age_{max}) for adapting the topology to the given environment. These experiments have resulted in a series of topologies represented in figures 4.15 to 4.17. The parameter space has been explored for the following values; a_T=0.5 to 0.9, h_T=0.01 to 1.0 and age_{max}=5 to 40. The figures show a set of 2D neural representations resulting from projecting the N-dimensional network onto the plane. These are also accompanied by the histograms of the metrics associated to each functional synapse relating every node to every behaviour. The figures are listed in an ascending number of nodes. These range from 2 to 40. Furthermore, these graphs are accompanied by a histogram underneath, representing the mean value of the fitting error of the affordance values obtained with the SOFM portrayed above. The mean and variance error of these graphs has been printed in every graph for clarity purposes (values are too small to draw error bars).

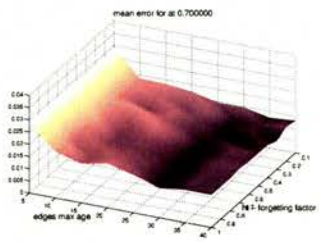
Figures 4.15 to 4.17 show the topology of the SOFM (green nodes and blue synapses) projected onto the data (red dots). As previously explained, the GWR used to cluster the sensory images is controlled by three main parameters: activity a_T, which controls the fitting accuracy of the network, habituation threshold h_T, which controls the time a node is allowed to reach its final destination and the age_{max}, which sets the maximum time a synapse can exist in the absence of excitation. When the habituation threshold h_T of a node is larger than the threshold h_T, the node does not appreciably move from its location. This drives the *growing* of the GWR in combination with the activity a_T. If the activity of the closest node is smaller than a_T the sensory space is considered to be underrepresented; and a new node is inserted. Furthermore, by increasing the habituation threshold h_T I am reducing the time given to any node to reach its final location, therefore also facilitating the insertion of new nodes. Consistently, when comparing the networks in figure 4.17; I can appreciate that in figure 4.17(b); a_T=0.9, h_T=0.5 and 4.17(c); a_T=0.9, h_T=0.7 the sensitivity to a variation of 0.2 of the habituation threshold means



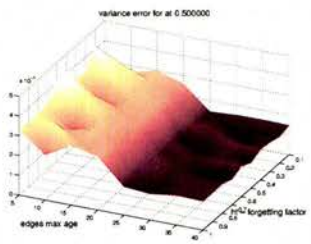
(a) Mean Fitting Error; $a_T = 0.5$



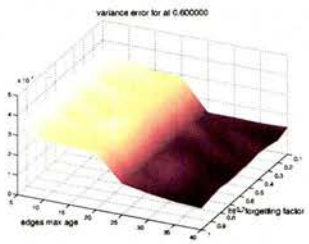
(b) Mean Fitting Error; $a_T = 0.6$



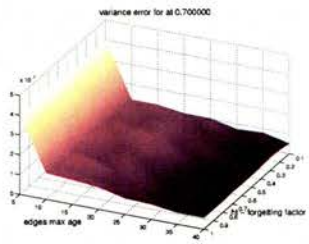
(c) Mean Fitting Error; $a_T = 0.7$



(d) Variance Fitting Error; $a_T = 0.5$

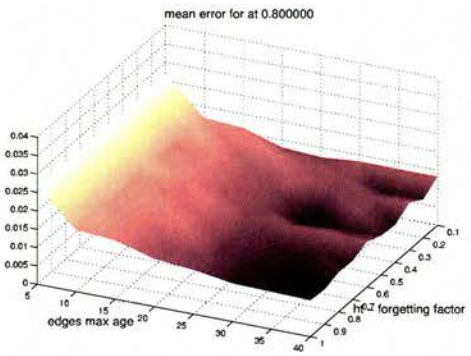


(e) Variance Fitting Error; $a_T = 0.6$

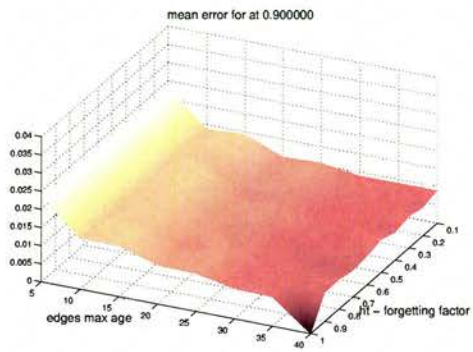


(f) Variance Fitting Error; $a_T = 0.7$

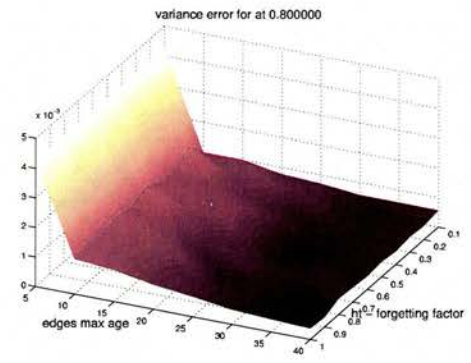
Figure 4.18: The surfaces represent the fitting error for the GWR algorithm for the case of the simple environment, mean and variance, top and bottom surfaces, respectively. The x and y axis stand for the h_T and the age_{max} parameters.



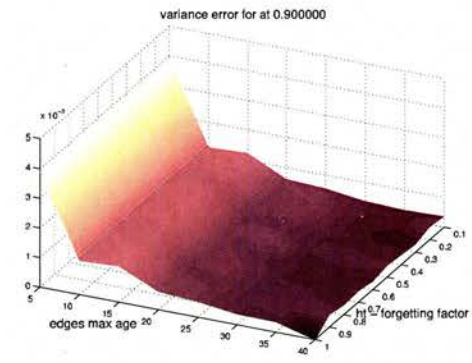
(a) Mean Fitting Error; $a_T = 0.8$



(b) Mean Fitting Error; $a_T = 0.9$



(c) Variance Fitting Error; $a_T = 0.8$



(d) Variance Fitting Error; $a_T = 0.9$

Figure 4.19: The surfaces represent the fitting error for the GWR algorithm for the case of the simple environment, mean and variance, top and bottom surfaces, respectively. The x and y axis stand for the h_T and the age_{max} parameters.

an increase of 5 nodes in the network. This is expected, since by increasing the habituation threshold h_T I am reducing the time given to any node to reach its final location, therefore facilitating the insertion of new nodes. Conversely, the *shrinking* of the network is mostly controlled by the age_{max} parameter. When synapses age over the age_{max} threshold, these are deleted together with the nodes that have no synapses. By comparing figure 4.15(b); a_T=0.5, h_T=0.2, age_{max}=5 to figure 4.15(c); a_T=0.5, h_T=0.2, age_{max}=20, it can be seen that the number of nodes increases as well. In general terms, the results demonstrate the general tendencies for each parameter of the network; namely, the higher the activity for each node a_T (the accuracy), the larger the number of nodes and consequently the more accurate is the topological map. Furthermore, the larger the forgetting factor h_T, the less time is given to the nodes to habituate, hence the more nodes it contains. The larger age_{max}, the more synapses the network contains.

In addition to these results, I have also considered it appropriate to calculate the *statistical stability* measured by the fitting error (equation 4.12) related for the affordance values averaged over 10 simulations (10,000 steps each). These are calculated according to the formulae 4.12 and 4.13 and are presented as surface graphs of figures from 4.18 to 4.19. The x-axis is the forgetting factor h_T, the y-axis the age_{max} and the z-axis the mean or the variance of fitting error for the topological network. In the sequence of figures it can be seen that the effect of an increasing a_T threshold between 0.5 and 0.9 diminishes the error of the affordance weights from 0.03 to 0.01. This error difference is significant, since from a behavioural perspective it means being able to correctly identify a node. Furthermore, it can be observed that the habituation factor h_T (x-axis) has little effect on the results and that there is a significant change in the fitting error for all graphs after a age_{max} larger than 30. After this value, the error dramatically diminishes to its minimum, the age rises over this value, and more nodes survive the deletion process; leaving the resulting GWR networks more densely populated.

Following these experiments, the criteria for a good parametrisation depends on the sensory pattern. However, the designer has to balance between a_T large enough to force the network to reach a certain accuracy and h_T low enough to allow the existing nodes in the network to habituate to their best fitting locations. Furthermore, this must be encompassed with age_{max} values large enough to allow this to happen before the synapses are deleted and the node is pruned. Given the experiments shown in this section, the sensory space would require a_T larger than or equal to 0.7, and age_{max} larger than 30. Final accuracy values have been gathered in figure 4.20, parametrised over the number of nodes of the resulting SOFM's. The conclusion is that the SOFM requires some parametrisation. However, for the given scenarios, where only two different objects are contained, two nodes seem to suffice for representing them (the error is only 0.02 in this case, cf. figure 4.20). As shown by this figure, there is not a clear advantage of using networks whose fitting error is larger. However, the next section will show that this is not the case when considering physiological metrics, since an agent endowed with an accurate

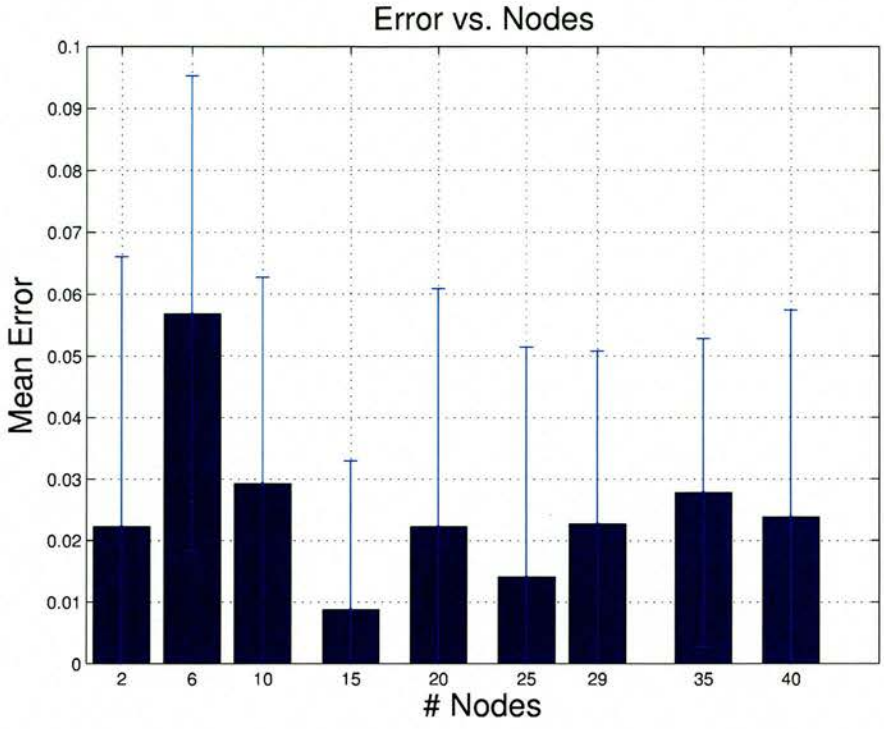


Figure 4.20: This histogram displays the mean fitting error for every node of the GWR networks, whose histograms have been previously shown

.

representation of the environment exhibits a more stable internal physiology.

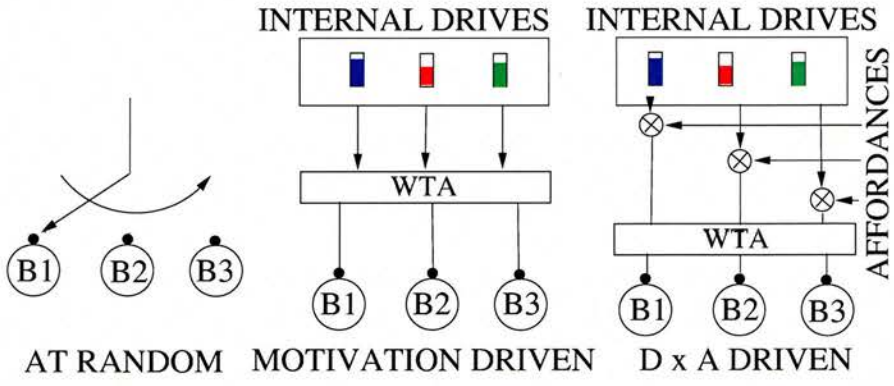
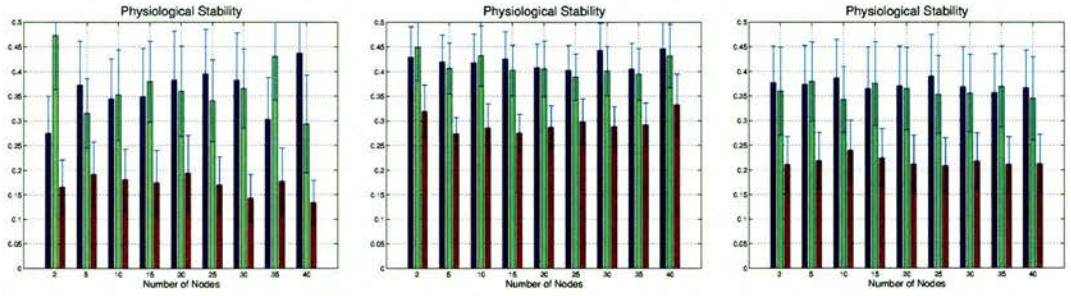


Figure 4.21: Strategies for Behaviour Selection. The behaviour the intensity of which is highest is dis-inhibited. Three procedures to calculate these intensities have been considered. From left to right, at random, motivation driven and $drive \times affordance$, respectively.

Physiological Measurements Finally, the analysis of this simple scenario is closed by a set of experiments addressing the measure of *physiological stability*, defined in equations 4.14 and 4.15. Previous experiments have considered the fitting error of the GWR network as a measure



(a) Physiological Stability; Random Selection

(b) Physiological Stability; Motivation Driven

(c) Physiological Stability; Multiplicative Formula

Figure 4.22: Physiological stability. The mean and variance value of the three drives (Blue=Hunger, Green=Tiredness, Red=Restlessness) averaged over 20 simulations are displayed. Simulations are parametrised after the number of nodes, ranging from 5 to 50 (x-axis), and after the policy to combine stimuli. Three different cases to combine stimuli have been tested; figures from left to right: random action selection, motivation-driven action selection, combined (multiplicative) action selection.

of quality to represent the environment. However, it is also important to consider the effect that one parametrisation or another may have at the behavioural level. The behavioural metrics have been introduced to this end, also parametrised after the different *mechanisms to arbitrate among behaviours*, namely random selection, motivation-driven (the behaviour whose related drive exhibits the highest value is selected) and a strategy using the multiplicative combination of stimuli to calculate the behavioural intensity ($affordance \times drive$), cf. figure 4.21. For the last two cases, the behaviour whose intensity is highest is selected. Figure 4.22 shows the stability for each homeostatic variable for the cases of (a) random action selection, (b) motivation-driven action selection and (c) multiplicative stimuli arbitration mechanism. These values are averaged over 10 simulations each. The best values correspond to the multiplicative formula, since they exhibit, in mean value, the most stable values for the three homeostatic variables. However, it can be seen that these values are not very far away from those obtained using random selection and that the latter are even better than those obtained using motivation driven policies. The explanation for this is that the multiplicative combination of stimuli is concurrently considering the external and the internal drives for action. This is sufficient to improve the stability of the agent's homeostatic variables. Furthermore, if only the internal motivations are considered (middle graph), this results in a lower stability because the external affordances are disregarded in the decision making. This highlights the notion that the way stimuli are combined for decision making is fundamental to gaining stability and therefore adaptation.

4.4.5 Learning Affordances in a Complex Environment

The experiments in the previous scenario have demonstrated that the GWR and the Hebbian learning algorithm proposed can be used to learn object affordances. Therefore, this same learning schema is proposed to be used in a more complex scenario. In this case, the new environment contains 10 octahedral objects, whose base-diameter size ranges from 0.01 to 0.1 in increments of 0.01.

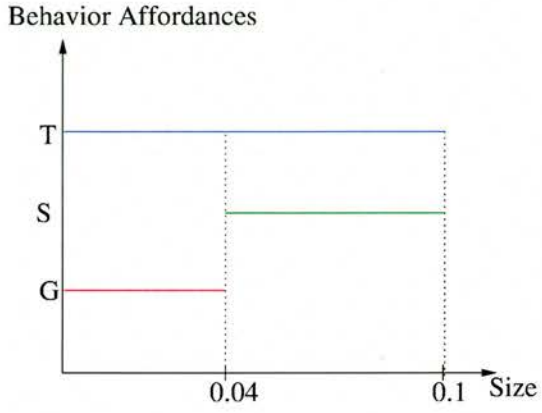


Figure 4.23: Abundant Distribution of Affordances. The line indicates the interval of sizes where that affordance is 1.0. G, S and T stand for grasping, shelter and touch, respectively.

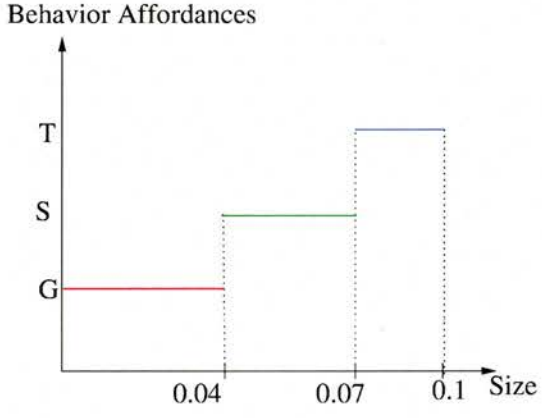
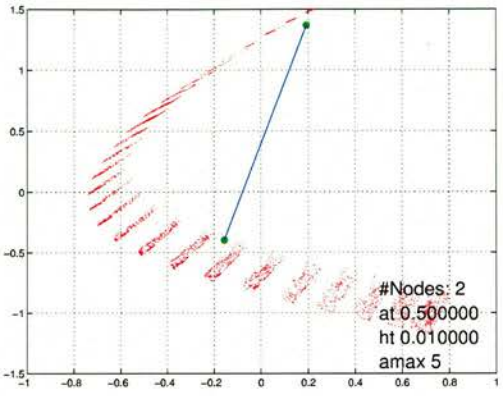
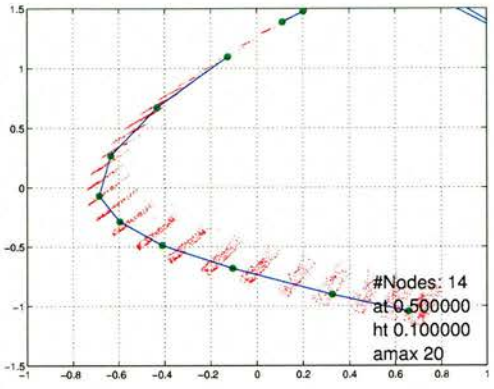


Figure 4.24: Scarce Distribution of Affordances. The line indicates the interval of sizes where that affordance is 1.0. G, S and T stand for grasping, shelter and touch, respectively.

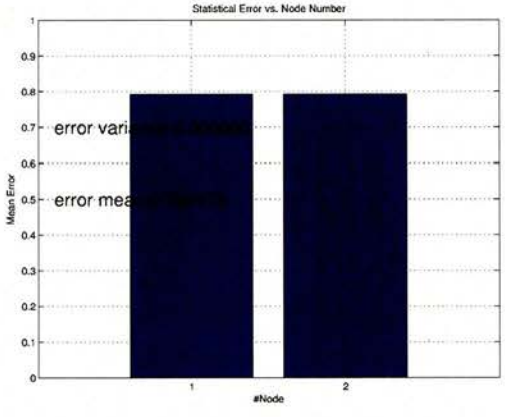
As designers, it has been considered appropriate to study two particular distributions of affordances: an *abundant* and a *scarce* distribution of affordances. The former leads to a scenario where most objects afford more than one behaviour to be executed; and the latter to a scenario where every object affords a single behaviour to be executed. These distributions are illustrated in figures 4.23 and 4.24, by showing the relationship between the size of an object and the behaviours they afford.



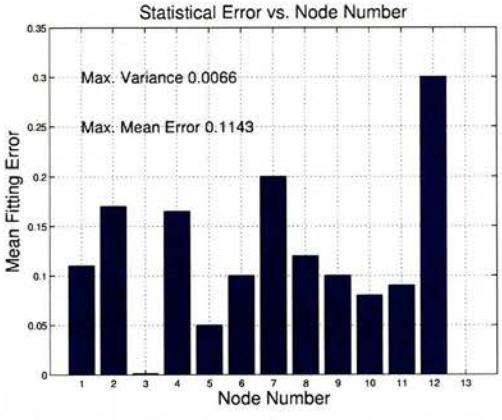
(a) 2N-GWR;$a_T = 0.5$, $h_T = 0.01$, $age_{max} = 5$



(b) 13N-GWR; $a_T = 0.5$, $h_T = 0.01$, $age_{max} = 20$

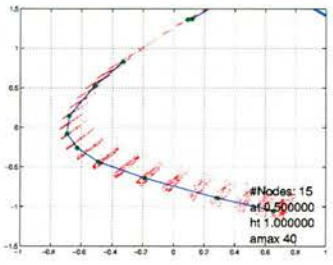


(c) 2N-GWR; Fitting Error.

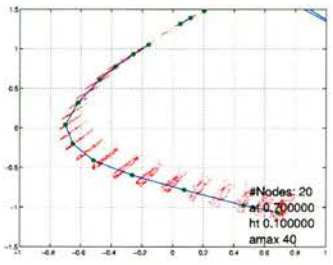


(d) 13N-GWR; Fitting Error.

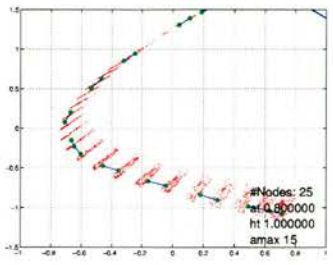
Figure 4.25: Top: 2-D Principal Component Analysis (PCA) of two GWR networks built for the case of the complex environment, with 2 and 24 nodes, left and right, respectively. The red dots are the samples used to train the GWR networks. Each GWR consists of a set of green nodes and blue synapses. Bottom: Histogram representing the mean fitting error for every node of the GWR network represented on top for every node of them.



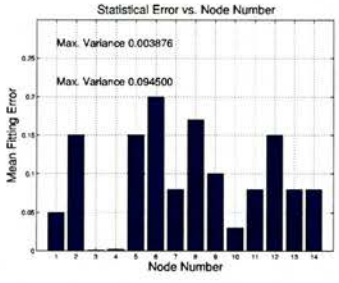
(a) 15N-GWR;$a_T = 0.5$, $h_T = 0.01$, $age_{max} = 5$



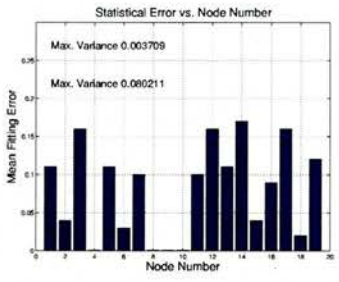
(b) 20N-GWR; $a_T = 0.5$, $h_T = 0.01$, $age_{max} = 20$



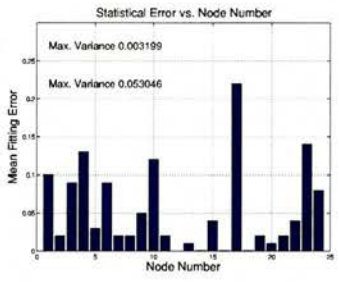
(c) 25N-GWR; $a_T = 0.5$, $h_T = 0.01$, $age_{max} = 20$



(d) 15N-GWR; Fitting Error.

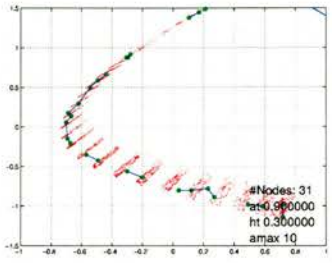


(e) 20N-GWR; Fitting Error.

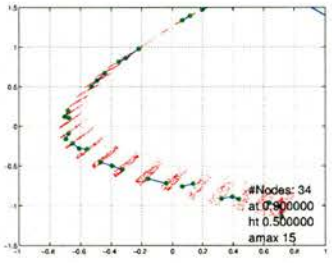


(f) 25N-GWR; Fitting Error.

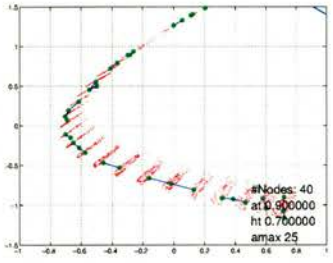
Figure 4.26: Top: 2-D Principal Component Analysis (PCA) of three GWR networks built for the case of the complex environment, with 15, 20 and 25 nodes, from left to right, respectively. The red dots are the samples used to train the GWR networks. Each GWR consists of a set of green nodes and blue synapses. Bottom: Histogram representing the mean fitting error for every node of the GWR network represented on top for every node of them.



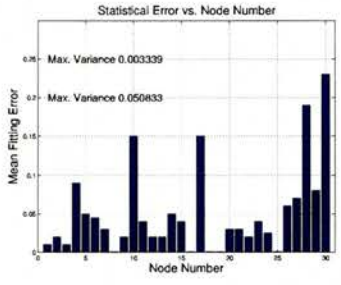
(a) 31N-GWR;$a_T = 0.5$, $h_T = 0.01$, $age_{max} = 5$



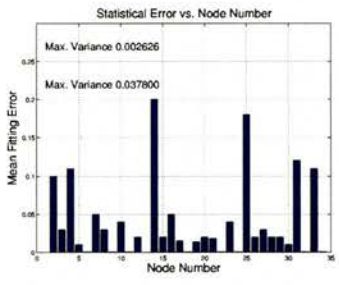
(b) 34N-GWR; $a_T = 0.5$, $h_T = 0.01$, $age_{max} = 20$



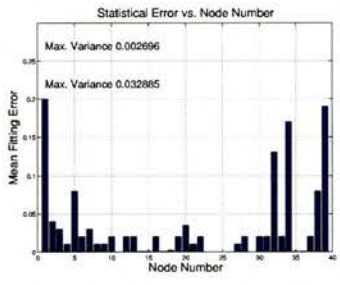
(c) 40N-GWR; $a_T = 0.5$, $h_T = 0.01$, $age_{max} = 20$



(d) 31N-GWR; Fitting Error.



(e) 34N-GWR; Fitting Error.



(f) 40N-GWR; Fitting Error.

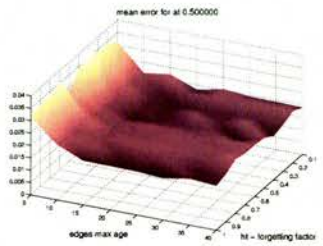
Figure 4.27: Top: 2-D Principal Component Analysis (PCA) of three GWR networks built for the case of the complex environment, with 31, 34 and 40 nodes, from left to right, respectively. The red dots are the samples used to train the GWR networks. Each GWR consists of a set of green nodes and blue synapses. Bottom: Histogram representing the mean fitting error for every node of the GWR network represented on top for every node of them.

As for the previous set of experiments, the first goal is to study the quality of the fitting of the sensory data that the GWR can provide. To this end, the distribution of affordances is irrelevant, since the topology of the GWR solely depends on the sensory data. This is independent of the function that any object affords to the agent. Therefore, the first (abundant) environment, characterised by a distribution of affordances, has been chosen at random to run the *first experiment set*, which addresses the relationship of the parameters to the *number of nodes* and the quality of the clustering of the sensory data. Furthermore, it also aims at analysing the *statistical stability* of the learnt affordance values. Figures 4.25 to 4.27 show the results obtained, accompanied underneath by the mean fitting error for each node of the network. Each figure shows on the top the 2D-PCA projection of the data (red dots) and the resulting GWR that clusters this sensory space according to the parameters specified in each graph. The statistical data is averaged over 10,000 steps of simulation —their overall mean error and variance over all nodes on each graph has been printed on every graph for clarity purposes (values are too small to draw error bars). The parameters used for simulation have been chosen between the following pairs of values in order to provide an appropriate insight on the behaviour of the GWR network. a_T between 0.5 and 0.9, h_T between 0.01 and 0.1, and age_{max} between 5 and 20.

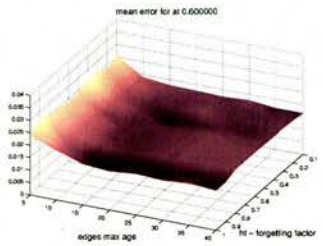
The results show that similarly to the previous experiments, the number of nodes for this scenario may be a vital parameter for the SOFM. In terms of fitting error, there is a dramatic difference between a SOFM of 2 nodes and SOFM of more than 10 nodes. This is also reflected in the mean error and variance of the weights linking the nodes and the behaviours. Furthermore, networks with more than 30 nodes fit the data accurately and result in a fitting error smaller than 0.05 on average.

This same error has been averaged over 20 simulations each and plotted in the 3D figures displayed in figures 4.28 to 4.29. Despite being a more complex environment, the effects observed are similar to those of a simple environment. The habituation threshold h_T (y-axis) has little effect. Conversely, the error sensibly diminishes when the activity of the nodes rises over a threshold value of a_T, since more nodes are necessary to cluster with a larger accuracy. Similarly, the error also exhibits an inflection point depending on age_{max} (x-axis) and is allowed to maintain sufficient nodes to cover the sensory space, as age_{max} increases, resulting in a larger accuracy and in a better fitting of the sensory space.

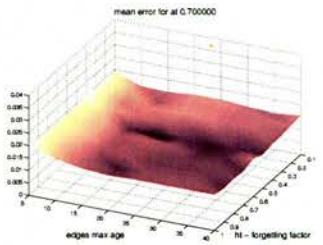
The last experiments for this complex environment pursue a double goal. Firstly, to *measure the required number of nodes for the SOFM in order to reach an optimal fitting error for the network*, before starting to overfit the given environment. Secondly, these experiments also intend *to measure the effect of the different strategies for behaviour arbitration.* As for the first environment, the strategies compared are: random selection, motivation-driven selection and selection according to the $affordance \times drives$ formula. Furthermore, a third parametrisation



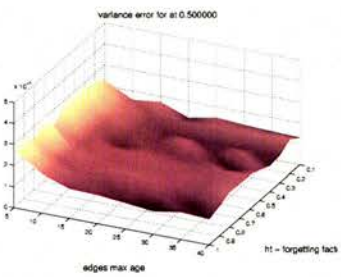
(a) Mean Fitting Error; $a_T = 0.5$



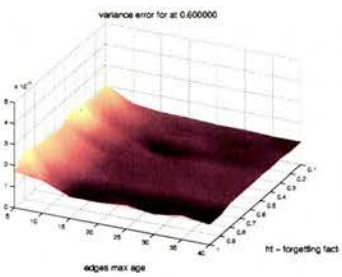
(b) Mean Fitting Error; $a_T = 0.6$



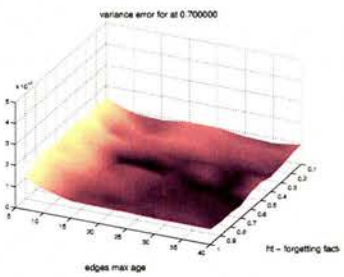
(c) Mean Fitting Error; $a_T = 0.7$



(d) Variance Fitting Error; $a_T = 0.5$

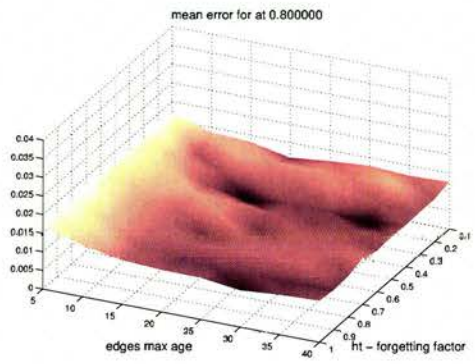


(e) Variance Fitting Error; $a_T = 0.6$

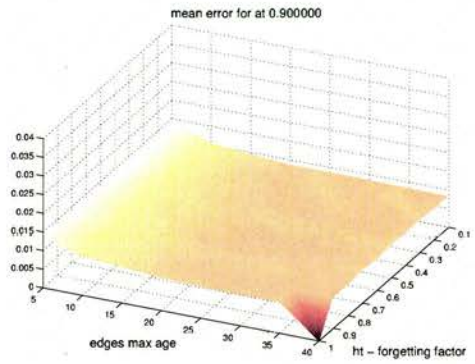


(f) Variance Fitting Error; $a_T = 0.7$

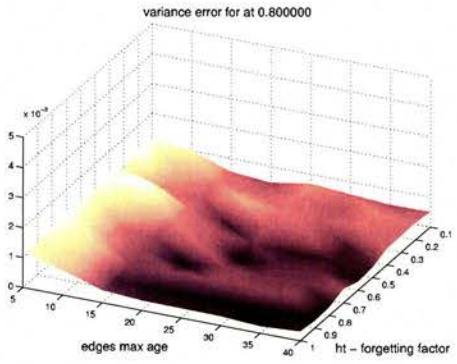
Figure 4.28: Mean and Variance of the fitting accuracy of the GWR networks, top and bottom surface graphs, respectively. The surfaces are parametrised after the activity of the GWR. It displays data for a_T=0.5, 0.6 and 0.7, from left to right, respectively. On x and y axes are plotted the age_{max} and h_T parameters.



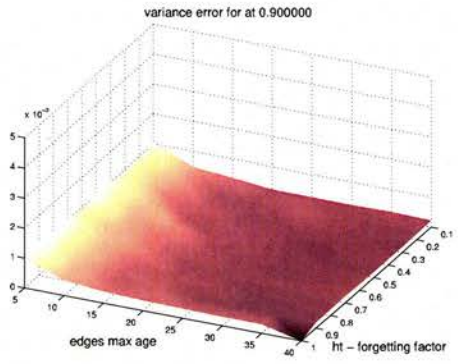
(a) Mean Fitting Error; $a_T = 0.8$



(b) Mean Fitting Error; $a_T = 0.9$



(c) Variance Fitting Error; $a_T = 0.8$



(d) Variance Fitting Error; $a_T = 0.9$

Figure 4.29: Mean and Variance of the fitting accuracy of the GWR networks, top and bottom surface graphs, respectively. The surfaces are parametrised after the activity of the GWR, values a_T=0.8 and 0.9, left and right, respectively. The x and y axis are plotted the age_{max} and h_T parameters.

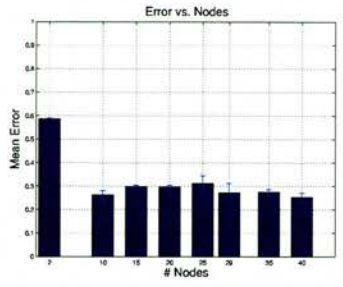
results from the types of affordance distributions considered, a *scarce* and an *abundant* distribution (cf. figures 4.23 and 4.24). The results, parametrised after the number of nodes of each environment and after the arbitration mechanism are shown in figures 4.30 to 4.32 for each distribution.

For both distributions, the top set of bar-graphs shows that for large number of nodes, 10 or larger, the accuracy of the SOFM to fit the sensory space is already reasonably asymptotic. The graphs in the centre show the values of physiological stability and overall comfort, calculated according to equations 4.14 and 4.15 for each individual drive of the agent.

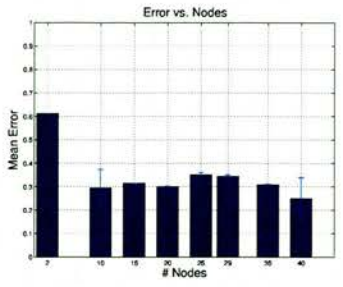
For the case of having an abundant distribution of affordances (D1), the values obtained are apparently very similar for either procedure to combine stimuli. However, the lower triad of graphs in figure 4.30 shows that the worst strategy for combining stimuli is the motivation-driven selection, since this disregards the object affordances for the selection. The winner among the three is again the multiplicative strategy. Despite its mean value being slightly higher than for the random selection, the error bars show a slightly smaller variance than for the random selection strategy; as a result the distribution is more stable.

Interestingly, experiments performed in the scarce distribution (cf. figure 4.24) demonstrate that both the number of nodes and the physiological stability are highly dependent on the influence of the external stimulus (the affordance). The centre and bottom graphs in figure 4.30 show that if the selection is at random or if it is motivation driven, the number of nodes of the GWR is not as important, since the selection of behaviour is performed disregarding the affordances in the scenario. However, the error bars on the right show that combining stimuli, both external and internal (affordances and drives), the number of nodes becomes fundamental to make decisions that lead towards a viable physiology. The reduction of physiological values when the number of nodes is highest demonstrates that the clustering and learning mechanism are working properly. Furthermore, these results demonstrate that it is necessary, for the scarce distribution, to use a SOFM with a large number of nodes in order to obtain accurate enough affordance values which, combined with the agent's internal drives, lead to the selection of behaviours which give physiologically stable values. Figure 4.32, another case of scarce distribution of affordances, confirms these results by showing similar results to those in figure 4.31. However, these results also highlight that the stability demonstrates a strong dependence on the rhythms of the agent's internal physiology. This will be addressed in chapter 6.

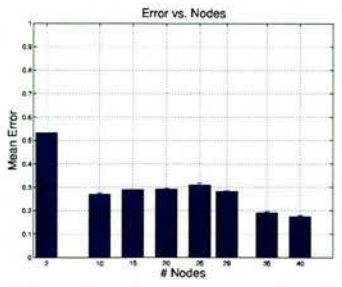
These results also highlight the fact that this mechanism is able to learn object affordances via relating the execution of a behaviour to the fluctuation that this provokes on the agent's internal milieu. This fluctuation is recorded by the hormonal responses of the agent, which control the synaptic weights relating the SOFM nodes to the behaviours; the affordance values of that particular scenario to that particular agent. These experiments have also highlighted that the fitting error quantifies the accuracy of the SOFM with respect to its sensory space. However,



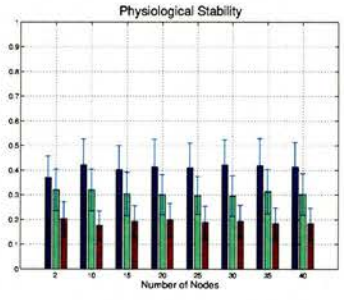
(a) Mean Fitting Error; $a_T = 0.5$



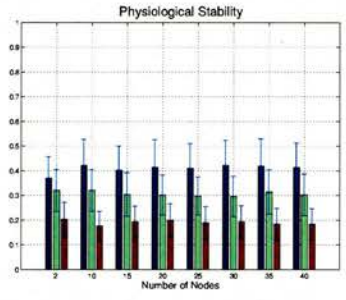
(b) Mean Fitting Error; $a_T = 0.6$



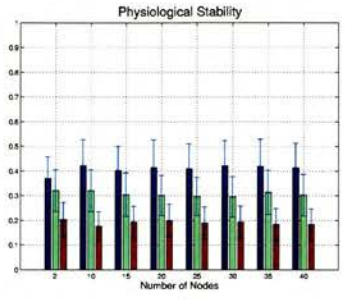
(c) Mean Fitting Error; $a_T = 0.7$



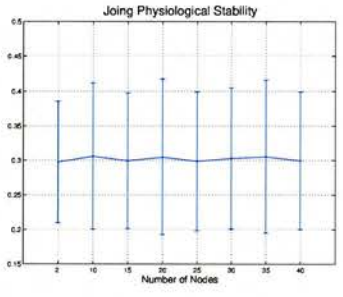
(d) m_i, σ_i; $a_T = 0.5$



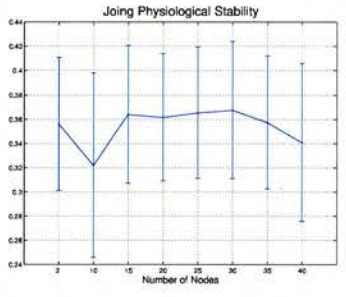
(e) m_i, σ_i; $a_T = 0.6$



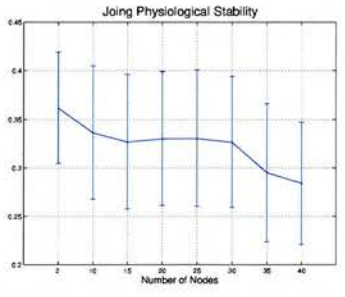
(f) m_i, σ_i; $a_T = 0.7$



(g) Viability; $a_T = 0.5$

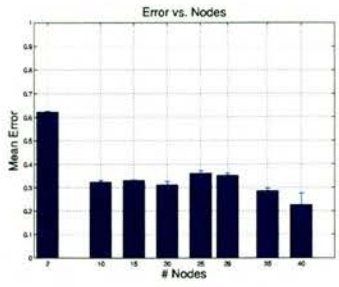


(h) Viability; $a_T = 0.6$

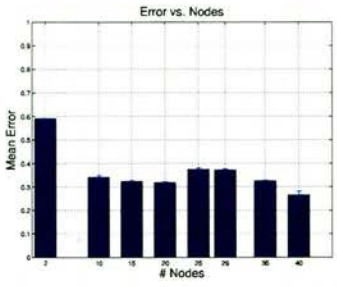


(i) Viability; $a_T = 0.7$

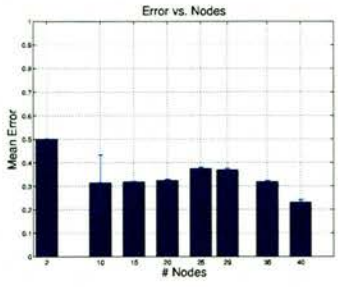
Figure 4.30: Physiological stability for the case of the *scarce* distribution of affordances, as specified in figure 4.24 for homeostatic variables decay constant $\tau = 10^{-4}$. The simulations have been parametrised after the number of nodes, from 2 to 40 (x-axis) and after the policy for the combination of stimuli; from left to right: random action selection, motivation-driven action selection, combined (multiplicative) action selection. The **top graphs** show the mean fitting error for the GWR networks used in these experiments. The **middle graphs** show the physiological values (mean m_i and variance σ_i of the physiological drives) for each drives of the agent. The **bottom graphs** show the viability indicators (physiological stability and overall comfort). These values have been obtained by averaging over 20 simulations. The colours, red, green and blue, correspond to the drives hunger, tiredness and restlessness, respectively.



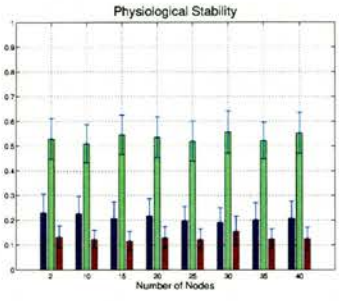
(a) Mean Fitting Error; $a_T = 0.5$



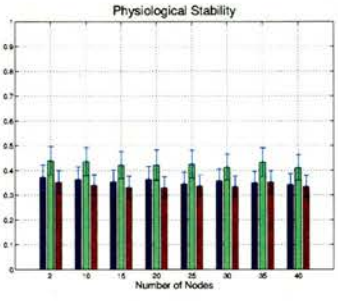
(b) Mean Fitting Error; $a_T = 0.6$



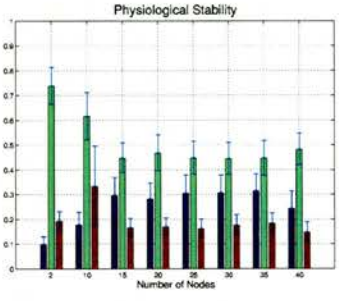
(c) Mean Fitting Error; $a_T = 0.7$



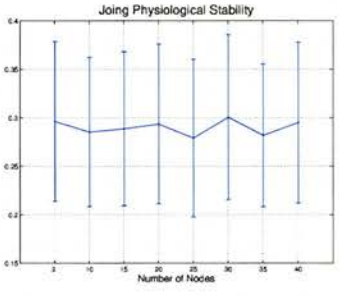
(d) m_i, σ_i; $a_T = 0.5$



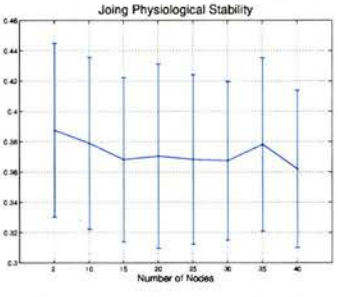
(e) m_i, σ_i; $a_T = 0.6$



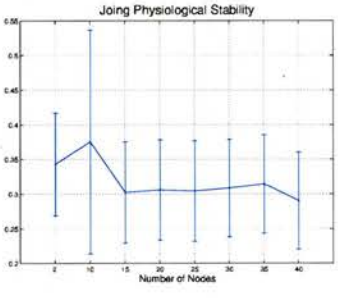
(f) m_i, σ_i; $a_T = 0.7$



(g) Viability; $a_T = 0.5$

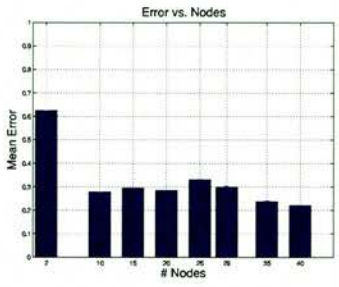


(h) Viability; $a_T = 0.6$

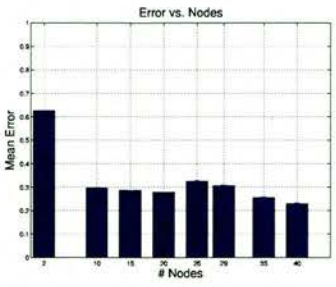


(i) Viability; $a_T = 0.7$

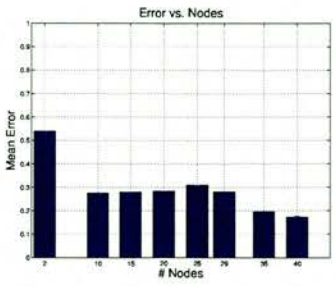
Figure 4.31: Physiological stability for the case of the *abundant* distribution of affordances, as specified in figure 4.23 for homeostatic variables decay constant $\tau = 10^{-3}$. The simulations have been parametrised after the number of nodes, from 2 to 40 (x-axis) and after the policy for the combination of stimuli; from left to right: random action selection, motivation-driven action selection, combined (multiplicative) action selection. The **top graphs** show the mean fitting error for the GWR networks used in these experiments. The **middle graphs** show the physiological values (mean m_i and variance σ_i of each individual drive of the agent). The **bottom graphs** show the viability indicators, physiological stability and overall comfort. These values have been obtained by averaging over 20 simulations. The colours, red, green and blue, correspond to the drives hunger, tiredness and restlessness, respectively.



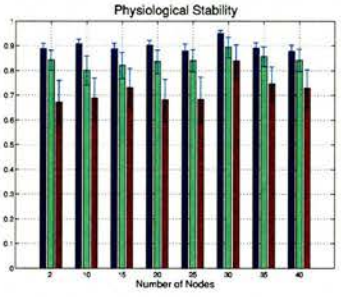
(a) Mean Fitting Error; $a_T = 0.5$



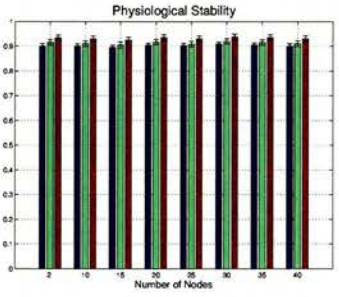
(b) Mean Fitting Error; $a_T = 0.6$



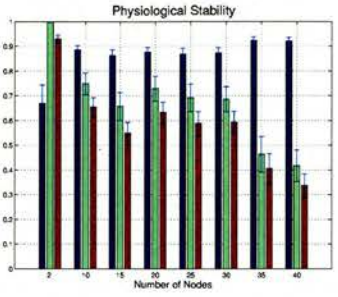
(c) Mean Fitting Error; $a_T = 0.7$



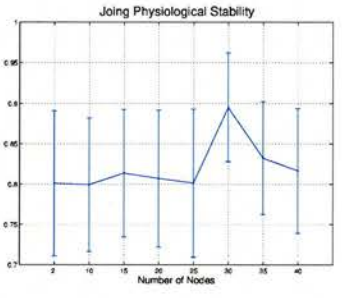
(d) m_i, σ_i; $a_T = 0.5$



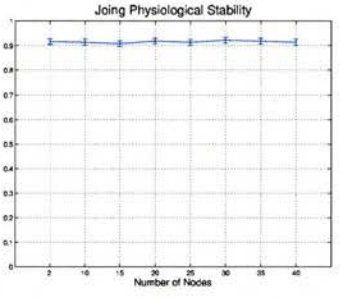
(e) m_i,σ_i; $a_T = 0.6$



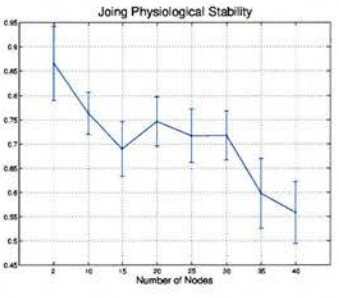
(f) m_i,σ_i; $a_T = 0.7$



(g) Viability; $a_T = 0.5$



(h) Viability; $a_T = 0.6$



(i) Viability; $a_T = 0.7$

Figure 4.32: Physiological stability for the case of the *scarce* distribution of affordances, as specified in figure 4.23 for homeostatic variables decay constant $\tau = 5 \times 10^{-4}$. The simulations have been parametrised after the number of nodes, from 2 to 40 (x-axis) and after the policy for the combination of stimuli; from left to right: random action selection, motivation-driven action selection, combined (multiplicative) action selection. The **top graphs** show the mean fitting error for the GWR networks used in these experiments. The **middle graphs** show the physiological values (mean m_i and variance σ_i of each individual drive of the agent). The **bottom graphs** show the viability indicators, physiological stability and overall comfort. These values have been obtained by averaging over 20 simulations. The colours, red, green and blue, correspond to the drives hunger, tiredness and restlessness, respectively.

as shown by the physiological metrics, this does not always mean that the agent will behave in an adaptive manner. In this respect, what really matters to the agent is being able to perform behaviours appropriately to maintain its physiological stability. This need of adaptiveness in physiological terms is the ultimate criterion to support the need and the quality of this metric, since only a behavioural demonstration, grounded in physiological measurements, can clearly demonstrate that the environment is perceived via a sufficiently accurate GWR network to the posed task.

4.5 Discussion

This chapter is about the ecological principle of adapting to the environment via learning object affordances. This principle is based on the existence of an animal in an environment as the result of their mutual interaction, at a developmental and at a genetic level.

The experiments introduced and discussed throughout this chapter focus on the developmental approach only; furthermore in the ability of learning the affordances offered by a set of objects to a simulated robot. It is argued that this ability endows an agent with the ability to adapt to a variety of scenarios in a straightforward manner. Furthermore, this approach can also shed light on the understanding of biological mechanisms of adaptation and would also provide an introductory assessment on whether these are applicable principles to mobile robotics.

Interaction with the environment throughout this chapter has been related to a formulation of reward as the mechanism of interaction between the agent and its environment, perceived as somato-sensory, kinaesthetic and sensory feedback affecting the agent. Modelling these phenomena in detail is a highly demanding task. However, for the goal at hand it is reasonable to simplify this interaction by a quantification of the effect of the execution of a behaviour on the agent's bodily dynamics. This formulation implies that the artificial agent exhibits an analogous structure to natural organisms; therefore, it consists of a set of internal resources (the homeostatic variables) and a set of drives signalling their status of deficit, normality or excess. Also for biological resemblance, the hormones satisfaction and frustration have been introduced to signal the sense of valence (Ackley and Littman, 1991) resulting from the execution of a behaviour (its beneficial or harmful consequences for the agent). Therefore, the agent has been assumed to be a part of a dynamical system also including the environment. This is conceptually novel from the perspective that the problem of relating perception to action (the problem addressed in mobile robotics since its most early stages) is now included in a larger framework, where self-organisation and learning are the mechanisms guiding adaptation at a developmental level. Biology has endowed animals with control mechanisms embodied in their neural systems. This has inspired the combination of the SOFM used to cluster the agent's sensory space and the Hebbian learning mechanism that builds up the agent's affordance pre-

diction substrate. Both mechanisms, synaptic plasticity and structural plasticity, respond in a concurrent fashion to the dynamics of interaction between the agent and its environment via modifying the agent's internal neural structure to better attain the goal of surviving. This has been related to maintaining the value of its drives as low as possible and making its internal dynamics most stable, according to Ashby's notion of viability (Ashby, 1965).

Both mechanisms, to cluster the agent's sensory space and to grow the functional synapses relating to the behaviours, are run concurrently within each simulation. This represents the dynamic nature of this adaptive process, where the inclusion or objects with novel affordances of a change in the distribution of affordances is tolerated by the agent if these changes are sufficiently slow. The results demonstrate that learning object affordances is possible with this formulation and that this learning enables the agents to maintain physiological stability in the scenarios considered. Furthermore, since the agent and its environment behave in a closed ecological loop, sensory input and behavioural responses are viewed as elements integrating elements configuring a dynamical system.

The system introduced in this chapter also exhibits some limitations. Firstly, clustering the sensory space and learning the synaptic weights relating each node to each behaviour have been mathematically specified as related processes. This may have a double interpretation; it may be argued that this has no conceptual effect and that the system is indeed based on the principle of ecology, since the interaction with the environment, hence the delivery of feedback by the environment is maintained. Nevertheless, it is correct that the biological resemblance is limited. It is likely that the agent's framework is implemented at a neural level by a single process integrating both the continuous processing of the sensory input (processing of the sensory flow) and the Hebbian learning of the functional synapses relating to the behaviour repertoire. This I have simulated by running both processes concurrently. Secondly, the source of feedback from the scenario has been modelled as a physiological fluctuation. However, at a biological level, the effect of an action may happen at different time-scales depending on the stimulus, e.g., the effect of being hit by a stone is nearly instantaneous, while that of eating a succulent meal usually happens at different time-scales; at the first bite the level of hunger is already modulated and does not stop until after it i sated. Finally, the assessment criterion used to attain the goal has assumed that behaviours for which the execution is physiologically beneficial are 'good' and that the others are 'bad'. It can be argued that this is common sense, since this is the same assumption guiding most actions in our life. However, this is also equivalent to hard-coding the sense of valence and therefore a restriction from a behavioural perspective, since addictive agents are discarded from the very beginning by design.

The behavioural observations obtained from the *last set of experiments* also confirm that beyond perception and triggering of behaviours, the strategy to arbitrate the selection of behaviours is vital to adapt to the environment. This introduces the problem addressed by the

following chapter: how does the agent learn behavioural patterns? Again this is guided by the ecological principle as a process that interacts with the learning processes related to the agent's perception: the learning of object affordances.

Summary

This chapter has introduced a framework to learn affordances based on the Gibsonian principle of ecological interaction between the agent and its environment. Affordances have been formulated as a neural structure relating nodes of a SOFM representing similar objects to behaviours within the agent's repertoire. The synapses of this structure are strengthened or weakened according to a Hebbian algorithm controlled by the agent's hormonal level. Each behaviour execution provokes a fluctuation in the internal physiology by increasing one homeostatic variable or another. This provokes the release of the hormones controlling the strengthening or weakening of the functional synapse relating the active node in the SOFM to the behaviour executed. This learning principle, relating interaction to physiology, is sufficient to learn object affordances in this context.

This perception-related process has been studied in so far as it can provide a sufficient representation for adapting to its environment at the developmental level. The learning principle is based on the relationship between the agent and its environment. The experiments have demonstrated that this is enough to learn object affordances. At a structural level the SOFM adds new nodes on demand of the sensory input. If an object with different features than those already represented is introduced in the scenario, a new node will be added when this is perceived. Furthermore, new functional synapses between the new node and the behaviours will be added to represent the affordances of the novel object. If changes are slow, the Hebbian algorithm computing the synaptic weights will be capable of re-calibrating in a continuous manner for a more efficient use of the resources. Conversely, abrupt changes in the environment may be too demanding for this algorithm and may lead to the 'death' of the animat. Physiological stability has been tested in scenarios containing small and large objects and distributions ranging from the scarce to the abundant. The performance has been assessed at a behavioural level by measuring the stability of the learnt functional weights, demonstrating the ability of the SOFM and of the Hebbian algorithm to continuously adapt to the given scenarios.

The use of affordances in the animal realm is not only controlled by perception but also by the intensity of the agent's internal drives. Affordances must therefore be tested in a larger framework of processes driving the agent's behaviour. I have performed a set of experiments to test different strategies for the combination of affordances and internal drives for decision making. The results show that the multiplicative formula $affordance \times drive$ produces behavioural patterns such that the agent's internal milieu is most stable. Furthermore, the differences in per-

formance obtained by using different strategies have also highlighted that the way of combining stimuli for decision making is critical for an efficient use of the affordances.

Affordances are learnt from the link between the selection and execution of a behaviour and the resulting physiological effect. I argue that this, the ecological link, must also be considered in the larger framework of adaptive processes, since ecology drives each adaptive process within the agent. This should in principle suggest that agents designed according to these general ecological views should adapt to their scenario better than those of agents designed under consideration of a single adaptive process, either perception or behaviour selection. Ecology drives each adaptive process of the agent, also the process of behaviour arbitration, which is addressed in the following chapter.

Chapter 5

The Actor-Critic learns Behavioural Patterns

The actor-critic algorithm has been postulated to drive the learning of stimulus-reward associations in Pavlovian contingencies (Schultz et al., 1993). In addition to this, the actor-critic has also been hypothesised to be involved in the learning of high level behavioural patterns (McClure et al., 2003; Houk et al., 1995) in higher vertebrates, such as mammals. These two hypotheses have been concurrently formulated in neuroscience; however, a model to conciliate both views, biological learning and behaviour selection, of application to robotics is still missing.

From the perspective of robotics, the use of reinforcement learning is far from being new, and likewise the use of an actor-critic (Sutton and Barto, 1981). Unlike other reinforcement learning algorithms, the actor-critic consists of two separate structures: the actor, to select behaviours, and the critic, to modify the behavioural patterns. This structure has been demonstrated to offer a lesser performance in terms of learning velocity than other algorithms where learning and selection are part of a single unit, e.g. Q-Learning (Sutton and Barto, 1998). Nevertheless, evolution in animals seems to have adopted an actor-critic as a solution to learn behavioural patterns, as supported by related studies of the role of the basal ganglia in mammals (McClure et al., 2003). However, several issues, including the role of extra-synaptic dopamine, still make its role difficult to understand. The learning hypothesis of Schultz (Schultz et al., 1993) and the action selection hypothesis of Redgrave (Redgrave et al., 1999) are not reconcilable at least due to the different roles hypothesised for extra-synaptic dopamine (DA). For the former, DA signals the error in the prediction of reward, for the latter it is the threshold to facilitate or hinder the selection of actions.

The model introduced in this chapter adheres to the view of Schultz and to the reinforcement learning view of the basal ganglia. However, it also aims at providing a justification of the possibility of the basal ganglia being a centralised action selector according to Houk's hypoth-

esis. This does not only consist of using the old actor-critic algorithm, but of defining a whole framework for its integration in a biologically inspired model. This model has been designed to this end; it consists of a motivational module and has integrated the perception system introduced in the preceding chapter. Therefore, if affordances can be viewed as a biologically inspired implementation of a reactive architecture, the learning of behavioural patterns introduced in this chapter can be viewed as related to higher cognitive functions by modulating behavioural responses. As an inspiration for this, learning is presented as a motivation-driven mechanism, grounded in the causality of interactions with the environment and their effect on the motivational state. Beyond the level of performance that these mechanisms may provide, they are intended to provide a demonstration of the aforementioned hypotheses and a suitable architecture to build autonomous agents with ideas inspired from neuroscience and ethology. Experiments have been performed in a range of scenarios in a simulated robotic platform addressing ethologically relevant situations.

5.1 Introduction

The issue of "knowing what to do next" has been addressed from different perspectives during the last decade (McClure et al., 2003; Dayan and Balleine, 2002; Avila-García and Cañamero, 2002; Cooper and Glasspool, 2002; Prescott, 2001; Spier and McFarland, 1996; Tyrrell, 1993). Nevertheless, a *grounded explanation of the effects, relating internal physiological dynamics, reward-based decision making and the involved perception processes is lacking in the literature*. This chapter introduces a view of these processes as elements of adaptation. This has been grounded on ideas of classical ethology, behavioural psychology and neuroscience and has led to a model that integrates motivation, behaviour selection, learning and perception in a biologically plausible manner.

Chapter 2 proposed a hierarchy of adaptive processes encompassing behaviour selection as the process that manages the changes of activity of the agent, and learning as the supervising process that modifies the patterns of selection. The interrelation between both has been implicitly suggested by previous implementations, although Bindra was the first to explicitly suggest a physiological and anatomical relationship between them: *"the effects on behaviour produced by reinforcement and motivation arise from a common set of neuro-psychological mechanisms, and the principle of reinforcement is a special case of the more fundamental principle of motivation"* (Bindra, 1969). This explicitly relates motivation, behaviour and reinforcement as processes sharing a common neuro-physiological substrate.

The concepts of motivation and reinforcement have been previously addressed and possess different nuances. *Motivation* was introduced in its different flavours by McDougall (1913); Freud (1940); Tinbergen (1951); Lorenz (1966). All of these definitions "share the idea of

a substance, capable of energising behaviour, held back in a container and subsequently released in action" (Hinde, 1960). According to this, motivation intrinsically relates the internal physiology to the behaviours, whose execution affects the intensity of their related motivations themselves. Therefore, motivation can be viewed as the modifier of behavioural tendencies via response instigation[1].

In a complementary manner, *reinforcement* events affect the future probability of choosing a certain behaviour, depending on the outcome of the interactions. Therefore, one can view response instigation as working forward in the direction from perception to action and that response reinforcement works backwards. However, both processes modify the future probability of choosing one behaviour over another. This mutual interaction, together with the use of a common neural substrate, suggests that both motivation and reinforcement can be measured with the same currency.

Related to this, the *common currency problem* (Redgrave et al., 1999; McFarland and Sibly, 1975) arises from the necessity to compare different *motivational states* in order to decide the appropriate behaviour to execute next. Hereby, it is argued that the common currency is reward. On the one hand, this would be consistent with the aforementioned notions of motivation and reinforcement, since reward can be viewed as the strengthening principle of the common neural substrate that instigates or discourages previously selected behavioural tendencies. Furthermore, decision making could then be conceived as a simple comparison among motivational states in terms of reward. The decision making (vanilla version), would then consist of choosing the behaviour whose related motivation is the most intense, since this is expected to lead to the highest reward. On the other hand, this definition of common currency is consistent with current neurological views of hypothetical roles for the basal ganglia (Houk et al., 1995).

In this respect, reward leads to reinforcement as a modulator of behavioural tendencies. *Learning* can be naturally integrated in this context if neuro-modulatory phenomena (Fellous, 2004; Fellous and Suri, 2003; Usher and Davelaar, 2002; Fellous, 2001; Hebb, 1949) are considered in the context of reinforcement comparison algorithms (Sutton and Barto, 1981). The reason for this is that these algorithms have been hypothesised to drive the learning of behaviour selection in circuits in the basal ganglia of vertebrates (Houk et al., 1995). Hence, if these hypotheses hold, striatal cells in the basal ganglia may be the "common circuitry for motivation and reinforcement" suggested by Bindra's hypothesis (Bindra, 1969).

However, traditional views on reinforcement, as shown in the model of Schultz et al. (1997); Suri and Schultz (1998); Houk et al. (1995); Schultz et al. (1993) seem to contradict models on pure behaviour selection hypothesised for the same parts of the basal ganglia (Gurney et al., 2001b,a; Redgrave et al., 1999). One of the *goals* of this chapter is to *to help*

[1] Biasing of a behavioural response.

to disambiguate both views: Schultz's view on learning and Redgrave's hypothesis on action selection. To this end, I have firstly decided to adhere to the view introduced by Schultz et al. (1993) on the role of dopamine and propose to study a set of related ethological phenomena. These can be tested at a behavioural level in order to demonstrate the coherence of this view via criteria analogous to those used in ethology. The main difference between both sets of models is the role of dopamine, which I assume to be the reinforcer (the error in the prediction of reward) that tunes tendencies in favour of one behavioural strategy or another (for Redgrave, dopamine only acts as a threshold that facilitates the disinhibition of a behaviour). In this light, a computational model has been built to address the process of learning to select among any sort of behaviours for a set of given environments. It has been tested in a simulated robotic architecture.

One of the fundamental ideas for the integration of both views originates in the psychological literature. Thereby it is straightforward to view *learning* as a step beyond pure behaviour selection. Behaviour selection is about changing activity and learning about modifying these patterns responding to changes in the environment. Three main elements are affected by learning: the *behavioural patterns*, *the perception of the world* (addressed in the previous chapter), and *the perception of the assessment* (addressed in the next chapter) itself. This chapter focuses on studying the underpinnings relating the availability and the distribution of affordances, the effect of the execution of a behaviour and the resulting behavioural patterns. Furthermore, from a more engineering perspective, I also aim at providing elements of design for long-term autonomous agents in dynamic scenarios.

The next section introduces the biological elements on which the model is based. This precedes a description of the elements used to build the behaviour selection and learning architecture. These are used to perform a series of experiments, addressing the integration of appetitive and consummatory behaviours in the competitive architecture for behaviour selection and learning. Furthermore, elements of design relating internal physiology dynamics are introduced in the next section.

5.2 Principia Biologica

I address the problem of choosing **what to do next** taking inspiration from biological systems. To that aim, the review in chapter 2 introduced the principles of learning prepared by Schultz et al. (1993) and of behaviour selection that Redgrave et al. (1999) derived from studies of the basal ganglia. The former argues that one of the roles of the basal ganglia is learning to relate stimuli to reward, and that dopamine exerts the role of signalling the error in the prediction of future reward. Conversely, the latter argues that the main role of the basal ganglia is the selection of actions. The role of dopamine in this case consists in facilitating the selection

by reducing (or increasing) its basal level (acting as a threshold for the activation of each behaviour).

Both hypotheses, *learning* and *behaviour selection*, have differing inspirations and only recently has it been suggested that both issues may be encompassed in the dynamics of a larger context (McClure et al., 2003). A model of learning based on Schultz's hypothesis has been proposed by Dayan and Montague (see Section 2.5). This is based on Schultz's hypothesis, which suggests that DA-cells signal effective reinforcement in the process of associating stimulus to reward. That is, DA represents the difference between the real and the expected reward given a certain stimulus. This is coherent with the experiments on macaques by Schultz (Suri and Schultz, 1998; Schultz et al., 1993), for which, after several trials training the macaque with the same stimulus-reward sequence, the habituation of the level of DA can be interpreted as a decrease of error in the prediction of reward for that particular stimulus.

Beyond the aforementioned interpretations, some further assumptions need to be made in order to integrate learning and behaviour selection in the same framework in a robotic platform. The contingency of learning observed in experiments with macaques is Pavlovian, i.e., when a stimulus is presented to the monkey, a reward will follow (or not). Nonetheless, the relevant observation is that the monkey has to perform *no action* to mediate its delivery. Unlike this, instrumental learning includes an action that needs to be executed after the stimulus is presented in order to deliver the reward. Hence, loosely speaking it is possible to view Pavlovian learning as a subset of instrumental learning, where the action is missing, or more formally, where the reward is mediated via the null action (in its absence). If this is correct, then the neural substrate of learning instrumental contingencies could also be the substrate of learning the Pavlovian. In the absence of further experimental data, it seems reasonable to assume that this is correct.

From a theoretical perspective, both hypothesised roles, that of reinforcement learning system and behaviour selection system, can be viewed as two particular sub-sets of a larger schema addressing learning and behaviour selection concurrently. In fact, it can be argued that to execute an action it is necessary to first select it. Furthermore, to that aim a strategy (which also has to be learnt) has to be applied to learn stimulus - action - reward relationships.

The model of action selection of Gurney, Prescott and Redgrave (Gurney et al., 2001b) is based on anatomical and physiological evidence and follows a bottom-up approach. Ethological, evolutionary and physiological studies suggest that the basal ganglia is effectively playing the role of a centralised action selector. I also support this view. However, Redgrave et al. (1999) argue that phasic dopamine response is too fast to be the predictor of reward when it occurs in response to novel stimuli, since it happens before the saccade to the stimuli is complete (assuming identification is necessary to signal reward). Conversely, recent evidence in neuroscience (Ross et al., 2001) suggests that "representations of future visual images may influence neural activity as if the saccade had already been executed (therefore, before the saccade), and

thus the dopamine response may anticipate the saccade" (Fellous and Suri, 2003). I argue that dopamine is signalling the reward of novelty, and that novelty can by itself be interpreted as a rewarding event. In the light of these arguments, it seems reasonable to assume the hypothesis of Schultz for the basal ganglia as an embodiment of an actor-critic reinforcement learning algorithm. This would include the thesis on learning, supported by Schultz's experiments, and partly also the thesis of Redgrave on behaviour selection. In this respect, the role of DA is assumed to signal not only effective reinforcement with regard to the stimulus, but also the error in the prediction of reward for the execution of one behaviour or another.

Furthermore, the ideas proposed by these principia can be related to the general theory on robotic architectures (Brooks, 1990). If affordances provide sufficient knowledge to build reactive architectures, it seems natural to extend these by integrating this knowledge in the actor-critic, since this would provide a superior layer to the architecture, which may extend its behavioural patterns and consequently provide adaptivity to a larger range of environments.

Based on these assumptions, I have built an architecture to integrate the actor-critic and the affordance-based perception system presented in chapter 4 in a single model. This is introduced in the next section.

5.3 Learning Motivational States

5.3.1 Introduction

Despite the number of definitions of motivation (Hinde, 1971), they all "share the idea of a substance, capable of energising behaviour, held back in a container and subsequently released in action" (Hinde, 1960). This principle was further formalised by McFarland and Sibly (1975), who proposed a representation of motivation in a formal state-space, assuming that "it is always possible to classify the behavioural repertoire of a species in such a way that the classes, which I call *activities*, are mutually exclusive". Hence, McFarland and Sibly were grounding the notion of motivation by relating internal physiology to behaviour, whose execution affects their related motivations. McFarland's categories were labelled as activities (throughout this thesis, they will be called *behaviours*), which can be uniquely determined if the causal factors, formalised as motivational state (the internal and external stimuli) are known. This is reflected in the following definition: "motivational state is the state value of all causal factors influencing a setup of functionally related behavioural patterns. The motivational state maps onto a *tendency* which is the strength of the behaviour in the competition for the final common path"[2](Toates and Jensen, 1990; McFarland and Sibly, 1975).

This *chapter describes ways of learning the relationship between each motivational state and the repertoire of behaviours, using the actor-critic based on the aforementioned princi-*

[2]Re-phrased from the cited papers.

ple of causality. According to the taxonomy of adaptive processes introduced in section 2.1, learning is the process of modification of behavioural patterns according to physiological (self-regulation) and environmental (availability, accessibility and distribution of resources) factors. Therefore, finding a solution to this problem will require:

- detecting the necessary elements from the environment to successfully execute the behaviour;
- knowing the appropriate behaviour to satisfy each internal homeostatic variable.

The previous chapter addressed the former point, formalised as *learning object affordances*. The latter is the core of this chapter. Figure 5.1 introduces the architecture for learning to select behaviours. The boxes for behaviour selection are depicted in black, in red are shown the modules for learning.

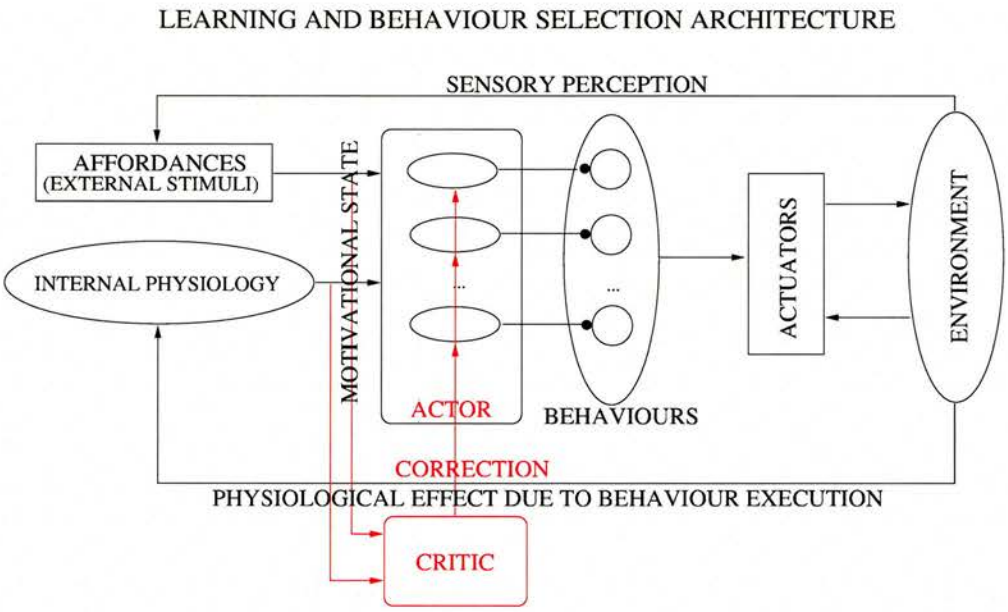


Figure 5.1: Architecture for Behaviour Selection and Learning. It consists of an internal physiology (homeostatic variables and internal drives), a module to learn affordances, a behaviour repertoire and the agent's actuators.

The *motivational state* consists of the set of motivations (Toates and Jensen, 1990) of the agent, these being a combination of external (affordances) and internal (drives) stimuli (Damasio, 2000; Toates and Jensen, 1990). The learning of affordances has been addressed in chapter 4. In this chapter it is assumed that the agent already knows the affordances of the objects in its environment (unless otherwise stated). For the case of motivation-driven agents, these are assumed to bias the execution of one behaviour over another, as shown by Avila-García and

Cañamero (2002). Hence, *learning to select behaviours* can be viewed as a search of the formulation that relates stimuli and drives (motivational state) to a sequence of behaviour executions that maintains the internal physiology within its viability zone. The process of evaluation of strategies is performed by the critic (represented by the red box in figure 5.1), which modifies the values of parameters of the policy functions (which calculate the motivational state) according to the error in the prediction of reward. The reward is due to the execution of a behaviour whose affordance is offered by the object the agent interacted with.

These principles set the framework for learning. However, a complete picture requires more definition: the *assessment criterion*. It is fundamental to know whether the execution of a behaviour is appropriate to the given motivational state; the problem is that there are as many correct criteria as definitions of goodness. However, it seems reasonable to assume that *behaviours whose execution leads to a higher physiological stability are beneficial*, conversely they are considered to be harmful. This derives from Ashby's notion of viability (keeping physiological variables within their viability zone). This assumption plays a dual role. On the one hand, it defines the necessary criterion of assessment for learning, on the other, it implicitly defines the sense of valency (the sense of goodness of an action) (Damoulas, 2004; Ackley and Littman, 1991) for that particular agent.

The *calculation of motivational states* has already been addressed by several authors. In particular, Spier and McFarland (1996) proposed the multiplicative constraint *Cue x Drive*[3]. This and other arithmetic formulations to calculate the intensity of each motivation have had some success in explaining some instances of animal behaviour. However, they have also failed to provide an explanation that relates physiological values to behavioural responses such as non-arousal (absence of stimulus should lead to a null motivational state) (McFarland, 1993) or null behaviour (expression of motivational leading to behaviour execution in the absence of stimulus). I argue that the formulation of the motivational state must be the result of a combination of external and internal stimuli. However, the combination of stimuli has so far been formulated according to multiplicative or additive formulae. Unlike this, I suggest that the combination of stimuli is probably non-linear and that it seems more advisable to let the interaction with the environment self-organise the behavioural responses on demand of the environment. To this end, I have introduced a set of non-linear estimators to combine the external and internal stimuli to be driven by interaction with the environment to learn these behavioural patterns. Therefore, I argue that no other formula must be *a-priori* specified in the learning process.

However, the calculation of motivational states does not explain the *selection process*. I have assumed that the behaviour to execute next is the one exhibiting the highest motivational

[3]Throughout the thesis, external stimuli are referred to in terms of affordances (and not cues). This is discussed in chapter 4.

value. It is however important to stress, that this is not the only criterion to select behaviours. Ours implicitly assumes that all motivations are measured with the same unit: *reward*, which makes the comparison possible. Although this selection criterion follows classic patterns, a restriction has been deliberately imposed: *the calculation of the motivational states and the selection process should lead to a sequence of behaviours stabilising the physiology of the agent and that maximises reward.* Unlike the case of arithmetic formulae, this principle provides a minimal restriction in the combination of stimuli. Furthermore, this has also been suggested to be a criterion of natural selection (Dawkins, 1976; Ashby, 1965), which provides a sufficient context for explaining a wide range of ethological phenomena.

A reward-based drive for behaviour selection has been proposed by different authors in machine learning (Sutton and Barto, 1998). However, this was not considered in neuroscience until recently as a possible explanation for behavioural patterns (Rolls, 2003; Schultz et al., 1993). The next section describes the implementation of these ideas and principles in our simulation environment.

5.3.2 Policy Learning Model

This subsection introduces a phenomenological model for learning behavioural patterns as an extension of a motivation-based behaviour selection architecture. The model consists of a framework for behaviour selection and learning based on the actor-critic reinforcement learning algorithm. The complete model is shown in figure 5.1, and consists of the following parts:

- The agent's *internal physiology* module integrates two sub-modules (cf. centre-left of figure 5.1) are two. Firstly, the homeostatic variables, which are abstractions of the agent's internal resources, and secondly the drives, which signal the status of deficit or excess of any homeostatic variable; their value is the difference between the optimal and the current value of their related variable.

- The agent's *external stimuli* (cf. top-left of figure 5.1). These are perceived as object *affordances*. These are quantified as weights of the synapses relating neural structure representing the sensory space (the SOFM) to the agent's behaviour repertoire. Refer to chapter 4 for more detail.

- The actor-critic algorithm makes decisions provoking transitions between different states in the Markovian state space. The state is uniquely determined by the values of the affordances and the values of the drives. In order to better understand the framework offered by the actor-critic algorithm, figure 5.1 shows a more detailed description of its different elements and their interactions. The state is represented in the centre-left. Based on this, the actor (top-centre) decides the behaviour to execute next. This piece

of information is also shared with the critic (bottom-centre), which guides the learning process on the grounds of the reward obtained due to behaviour execution.

- The behaviour repertoire is represented on the centre-right of figure 5.1.

Learning Framework As stated in the previous section, the calculation of the motivational state is not specified by an arithmetic formula. Instead, the actor-critic learns, via interaction with the environment, to compute the motivational states leading to an internal physiological stability. To do this, a set of *preference functions* calculates the intensity of the motivation for each behaviour; the behaviour related to the highest motivation is executed next. Therefore, the performance in terms of learning will be conditioned by the performance in selecting behaviours, which depends itself on the correctness of the prediction of reward for each behaviour. The predictions are updated after the execution of every behaviour by correcting the weights of the network that predicts the motivational intensity of that behaviour with the value of effective reinforcement, i.e., $\delta(t)$, the difference between the real and the expected reward.

To reach the correct policies for every state depends on two different conditions: firstly, to satisfy the conditions of convergence for the actor critic and secondly, to have an appropriate definition of reward. This definition should be a function of the effect of interactions on the agent's internal physiology. However, there is a series of considerations with regard to its biological plausibility. These are introduced below. On the one hand Houk et al. (1995); Schultz et al. (1993) proposed that the basal ganglia was encoding stimulus-response for learning for Pavlovian contingencies. On the other, it has been suggested that both Pavlovian and instrumental learning partly share the same neural substrate, since both seem to learn in a stimulus-response manner (McClure et al., 2003), as their only difference is that in instrumental learning the delivery of reward is mediated by the execution of an action. Therefore, it seems reasonable to extend a model that was originally proposed for Pavlovian learning, which also considers the selection of the behaviour, to instrumental learning: the actor-critic.

This algorithm is a derivative of the originally proposed *reinforcement comparison* algorithms (Sutton and Barto, 1981). According to Sutton and Barto, the reinforcement learning problem is "meant to be a straightforward framing of the problem of learning from interaction to achieve the goal" (Sutton and Barto, 1998), pp. 51. Viewed from a machine learning perspective, the problem includes three elements: the goal, the states among which the agent will have to navigate and the way the reward (assessment) is introduced in the system. In a straightforward manner, the states have been defined as a combination of internal physiological states and the affordances of the objects encountered, and reaching the goal as leading the agent's physiological space to its optimal zone. The definition of reward is addressed next.

Defining Reward Definitions of reward have often been misleading, since different authors mentioned reward when they refer to different phenomena. I define reward as *an affective qualification of the effect of a behaviour on the agent's internal physiology*, hence the feeling (Damasio, 2000) of the effect that particular agent. Reinforcement is the strengthening or weakening happening at a synaptic level.

The learning of behavioural patterns is based on the relationship between reward and the agent's internal physiological stability (see equation 5.1).

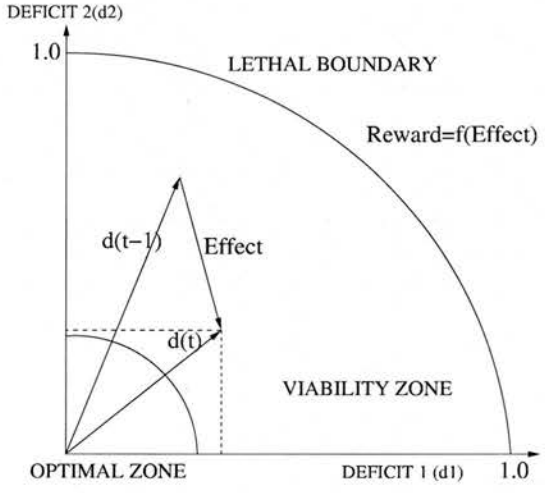


Figure 5.2: Definition of Reward. The x and y axes represent two deficits of the agent's internal physiological state. The deficits are represented by the vectors $d(t-1)$ and $d(t)$, before and after the execution of a behaviour, respectively. The effect of executing a behaviour is represented by the vector $Effect$.

The reward $r(t)$ is the difference of the inverse of the square of the norms of the vectors in figure 5.2, where $\bar{d}(t-1)$ and $\bar{d}(t)$ are the vectors representing the initial and final states, respectively, in the agent's physiological space.

$$r(t) = \frac{1}{\|\bar{d}(t)\|^2} - \frac{1}{\|\bar{d}(t-1)\|^2} \tag{5.1}$$

In addition to respecting the overall goal of reaching and maintaining the optimal of the agent's internal milieu (Ashby, 1965), this definition of reward has further consequences. Firstly, it implicitly relates to the definition of the feeling of what is good and bad for the agent, hence the sense of valence (Ackley and Littman, 1991) for this agent. Secondly, the non-linearity of the formula of reward is biologically consistent. If the execution of the behaviour leads the agent's physiological state to the origin ($d(t)$ is close to the origin), the formula will deliver more reward than if the effect leads the physiological vector $\bar{d}(t-1)$ further away from the origin. Hence, it is more rewarding to reach the stability zone than it is to just compensate

an urging internal need. However, the formula will deliver a negative value (interpreted as a punishment) if the behaviour executed tried to affect a related homeostatic variable.

From a conceptual perspective, it is important to note that reward only refers to positive contributions and that punishment refers to its negative counterpart. Lastly, this definition of reward is also defining the fitness of the individual (Dawkins, 1976), although this only has an effect at a developmental level for a single agent.

Cycle of Execution In most reinforcement learning algorithms, the selection of actions and the learning from the environment are integrated in the same module, e.g. Q-Values (Watkins, 1989). For the actor-critic module, these are deliberately separated. The actor is responsible for the measurement of the motivational state and for selection of the behaviour to execute next. The critic provides the necessary feedback for improving the policies for behaviour selection (cf. figure 5.3). On the one hand, this facilitates the selection of actions, since behaviour selection results from a comparison among the preference values (motivation values), on the other, "they [agents] can learn an explicitly stochastic policy; that is, they can learn the optimal probabilities of selecting various actions" (Sutton and Barto, 1998).

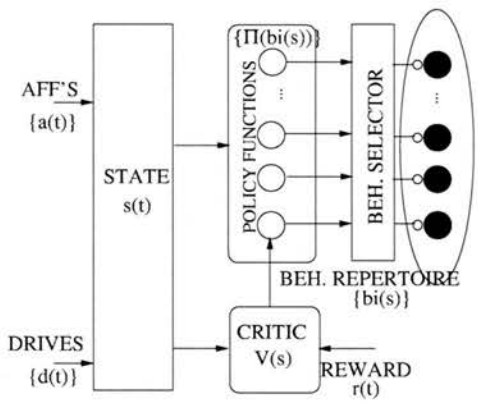


Figure 5.3: The Reinforcement Learning Schema consists of an Actor-Critic reinforcement learning schema. Its state consists of the set of perceived affordances and of the values of the agent's internal drives. This information is used to calculate the policy values associated to that state to make decisions. The critic provides the feedback to update the policies on the basis of the reward provided by the execution of each behaviour.

Hence, the *learning algorithm has two phases.* First phase: the *calculation of the motivational state* as a function of the physiological state and the selection of the behaviour to execute next. The physiological state $\underline{s}(t)$ (see equation 5.2) is a vector of real numbers, ranging between 0.0 and 1.0 and representing the drives $d(t)$ and the affordances $a(t)$. There are N homeostatic variables (hence drives) and L affordances. The mapping between the physiological and the motivational states is performed by feed-forward neural networks (non-linear

estimators); there is one network doing this mapping for each behaviour. The physiological state is updated, by perceiving the affordances of the object nearby and by reading the instantaneous values of the internal drives. Then the networks associated to each behaviour calculate the intensity of each motivation to execute each of them. The behaviour selection follows a greedy policy; 80% of the time the behaviour whose related motivation exhibits the highest intensity value — or policy value — (Π_{b_i} is the policy function, see equation 5.3) is selected for execution. The rest of the time a random behaviour is selected to explore the state space.

$$\underline{s}(t) = \{\{d(t)\}_i, \{a(t)\}_j\}, i \in \{1..N\}, j \in \{1..L\} \tag{5.2}$$

$$b(\underline{s}(t)) = argmax\, \Pi_{b_i}(\underline{s}(t)) \tag{5.3}$$

Second phase: the *learning phase*, which occurs after the selected behaviour has been selected and executed. The reward obtained via the execution of the behaviour, $r(t)$, is compared with the prediction of the critic. The difference between both is the error in the prediction of reward $\delta(t)$, see equation 5.4. γ is the discount factor ($0 < \gamma < 1$), specifying the influence of past states on the current state.

$$\delta(t) = r(t) - \gamma V(\underline{s}(t)) + V(\underline{s}(t-1)) \tag{5.4}$$

The critic estimates the value $V(\underline{s}(t))$ of the state $\underline{s}(t)$, in other words the cumulative value reward resulting from the successive interactions with the environment, from an initial state until the goal (the viability zone) is reached. The value function is calculated by a feed-forward neural network.

Therefore, equation 5.4 shows the difference between the real reward and the predicted reward: $\delta(t)$. This scalar is on the one hand used to update the weights of the critic for refining the prediction of reward for the next estimation via back-propagation (Rumelhart et al., 1986). This algorithm minimises the Minimum Square Eror (MSE) ($E = 1/2 \sum_{j=1}^{L} (d_j - a_j)^2$), where d_j are the desired output values and a_j the real outputs values, L is the number of units at the output layer. The goal is to minimise this error; hence the error is diminished at every step by gradient descent according to

$$\omega_{ij_t} = \omega_{ij_{t-1}} - \eta \frac{\partial E}{\partial \omega_{ij}}, \tag{5.5}$$

where ω_{ij} is the weight of the synapse connecting unit j at the middle layer to unit i at the output layer (or j at the middle layer and i at the input layer). Depending on whether the weight connects the output to the middle layer or the middle to the input layer, the expression of the error to backpropagate will be different. For the former case, the expression to update the error

i

$$\frac{\partial E}{\partial \omega_{ij}} = -(d_m - a_m) a_i, \tag{5.6}$$

where a_m is the real output and d_m is the desired output. For the latter case, the expression of the error is a little more complicated, and responds to

$$\frac{\partial E}{\partial \omega_{mi}} = a_i g(s_m)(1 - g(s_m)) \sum_{k=1}^{K} \omega_{km}(d_k - a_k), \tag{5.7}$$

where ω_{mi} is the weight connecting the i input unit to the m middle layer unit, a_i is the value of input i and $g(s_m)$ (g is the sigmoid function) is the output of unit m, ω_{km} is the weight connecting the middle layer unit m to the output unit k. d_k is the desired output value and a_k the real output layer value at the output unit k.

Equations 5.6 and 5.7 describe the process of updating the weights of the output and middle layers for the critic and for every behaviour network included in the actor. I need to notice two main differences with respect to the usual algorithm. The first one is that the output layer has a single unit, since the prediction of reward for the critic and the policy functions of the actor are scalars. Therefore, the summation over k units is reduced to a single term. The second difference is that for every one of these cases, the output layer is linear, since the policy functions and the error in the prediction of reward are unbounded.

The same process of weight updating is performed for the networks estimating the policy function for every behaviour. From a biological perspective, these two updates model the learning guided by the dopamine signal, the error in the prediction of reward in an instrumental contingency. Thus $\delta(t)$ can be linked to the dopamine signal, the error in the prediction of reward observed for Pavlovian learning (Schultz et al., 1993). Both selection and learning operations are repeated until convergence. The convergence is facilitated by a positive reward ($r(t)$) when the effect of the chosen behaviour diminishes the drives of the agent, conversely for any other case.

It is argued that the framework introduced in this section, together with the aforementioned ideas on motivation and reward are a sufficient context for integrating learning and behaviour selection in a biologically plausible manner. This argument will be addressed in the following sections by testing the learnt policies in a set of scenarios, where the availability, accessibility and distribution of resources has been distributed in a biologically plausible manner.

5.3.3 Experimental Setup

The principle for learning appropriate behavioural patterns consists of grounding the interaction with the environment to its physiological effect. This relates to the reward obtained by that execution, which depends on the dynamics of the internal physiology. For testing purposes, a set of different environments have been engineered, varying the availability, accessibility and distribution of resources. It is expected that any change in these variables will be reflected in appropriate changes in the behavioural patterns in order to adapt to each scenario.

The *model* is composed of three homeostatic variables and three drives. These are structured in modules, as introduced in section 4.1. Three *homeostatic variables* have been labelled as Nutrition, Stamina and Boredom. Each of them has a related *drive*, namely hunger, tiredness and restlessness, respectively. Their value is computed as the difference between the optimal and the current level of the related homeostatic variable.

The *behaviours* are coarse grained and integrate a set of actions. The behaviours used for the experiments are the following: to grasp, to drink, to shelter, to touch and to avoid. They are consummatory, except the last one, which is appetitive[4]. One of the constraints of the agent is that only one behaviour can access its actuators at a time. The rest of the time the behaviour must be inhibited. The task of the actor-critic model is to *learn to associate each state to the suitable behaviour*, the state being the combination of external affordances and internal drives, cf. equation 4.2.

Metrics The overall goal is to select behaviours appropriately to maintain the agent's internal physiology within its viability zone. Three sets of metrics are introduced in order to assess the performance of the learning algorithm.

Firstly, it has been considered appropriate to measure the time it takes to learn the behavioural patterns to respond to the environment. To this end, the *number of steps to reach stability* has been sampled every 2000 decisions until a stable value is reached (the error in the prediction of reward δ is smaller than ε). The time to reach this is the time of convergence for the optimal selection pattern, given the particular scenario and the conditions of the algorithm.

Secondly, the goal of the algorithm is to bring stability to the agent's internal physiology. To this end, two *viability indicators: physiological stability* and *overall comfort* (refer to equations 4.14 and 4.15, respectively), are measured at intervals of 2000 decisions; until convergence is reached. Physiological Stability measures the mean value of the deficits, and overall comfort measures how much these vary. The optimal value is 0 for both cases. The evolution of these indicators over time displays the effect of the different behaviour selection strategies in physiological terms, starting with the initial policy (at random) and concluding with the final (reward optimal) policy.

Lastly, for ethologists, it is fundamental to observe behavioural responses. I have gone one step further than this by relating the behavioural responses to their physiological state. To this end, *behavioural and physiological cycles*[5] have been drawn with a dual goal; on the one hand to depict the relationship between the physiological state and the learnt behavioural patterns (considered appropriate to the situation described by the state); on the other to have an element

[4] A behaviour is considered consummatory if its execution provides immediate reward to the agent. Otherwise it is considered appetitive.

[5] A behavioural cycle is the sequence of behaviours chosen and executed by the agent to compensate a physiological state. It ends when every deficit has been covered.

of behavioural analysis analogous to ethological cycles[6] (McFarland and Spier, 1997).

5.3.4 Relating Physiology, Behaviours and the Environment

I firstly intend to test *the capacity of the proposed algorithm to learn to respond to the current physiological state with the right behaviour* by interacting with its environment. To this end, an environment consisting of a set of objects has been engineered. The encounter with objects in the environment has been simulated in order to speed up the experiments. Objects are encountered on a random basis. These simulated environments have been characterised according to the distribution of affordances of the objects as a function of their physical features. In this section, the goal is to test the learning algorithm in two environments D1 and D2, D1 is an environment where every object affords every behaviour to be performed. In distribution D2 objects whose size is smaller than 0.04 afford grasping. Objects larger than that size afford to shelter; all objects afford to be touched (cf. figure 5.4). This distribution is such that 50% of the objects afford shelter and 50% afford grasping. 100% of the objects afford to be touched. Hence, the availability of resources is diminished with respect to D1.

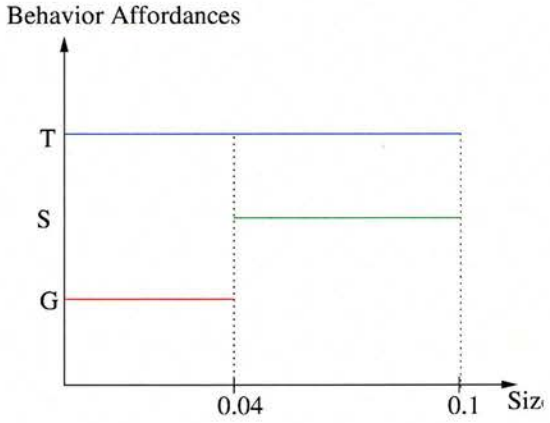


Figure 5.4: D2 Distribution of affordances. The line indicates the interval of sizes where that affordance is 1.0. G, S and T stand for grasping, shelter and touch, respectively. Object sizes range from 0.0 to 0.1.

The *learning procedure* consists of the following steps:

1. The agent encounters an object, whose affordances are known to the agent.
2. The policy values (motivational state) are calculated for each behaviour given the current state (the values of the affordances of the closest object and of the drives).
3. The behaviour that maximises the expectation of reward (whose motivational state exhibits the highest value) is selected and executed 80% of the time. Selection is at random the rest of the time.

[6]Ethological cycles are sequences of behaviour selection executions aimed at compensating certain needs.

4. If the execution is successful, it has a compensatory effect on one or more homeostatic variables; conversely, there is no effect.

5. The real reward is calculated as a function of this effect.

6. The critic performs the correction for its prediction of reward according to equation 5.4, and the actor corrects its preferences for the behaviour selected and executed, given that state.

These simulations address the following **issues**:

- The **learning of the correspondence between the drives and the behaviours**. This has to be evaluated via *cycles of behaviour execution* and *a metric of effectiveness*, see equation 5.8.

- The **characterisation of the learning process**. The metrics for these are: the *mean number of decisions required to reach the goal from a random initial point*, the *necessary number of decisions to reach the shortest cycle* and the *final stable values of physiological stability and overall comfort.*

Learning Performance and Physiological Stability The experiment aims to demonstrate the capacity of the architecture above to relate the dynamics of the internal physiology, characterised by the decay of the homeostatic variables ($\tau = 10^{-3}$), to the execution of behaviours for different scenarios. These scenarios have been parametrised after the distribution of affordances, D1 and D2 in this section, and the experiments aim at characterising the final behavioural patterns for both distributions. The experiments take place in the simulated environment described in section 5.3.4.

Graphs in figure 5.5 show the evolution of the length of the cycle of behaviour executions, averaged over 20 simulations. The cycle[7] starts at a random value in the agent's physiological space and ends in the optimal zone of the agent's physiology (where deficits are close to zero). The changes are due to the execution of behaviours and are used as an evaluation procedure. These graphs show that the mean *number of decisions* required to reach the optimal physiological zone decreases for both affordance distributions (top left graphs, red and blue curves, respectively) until a stationary value is reached (about 11 and 17 for cases D1 and D2, respectively). Analogously for the statistical metric of variance, cf. top right graph in figure 5.5. Furthermore, ca. 16×10^4 iterations (80 time steps) are necessary to reach the stationary value.

Similarly, the *viability indicators*, physiological stability and the overall comfort averaged over 20 simulations each (bottom graphs, left and right, respectively), show that stability and

[7]The name cycle is assigned by analogy with the notion of ethological cycle.

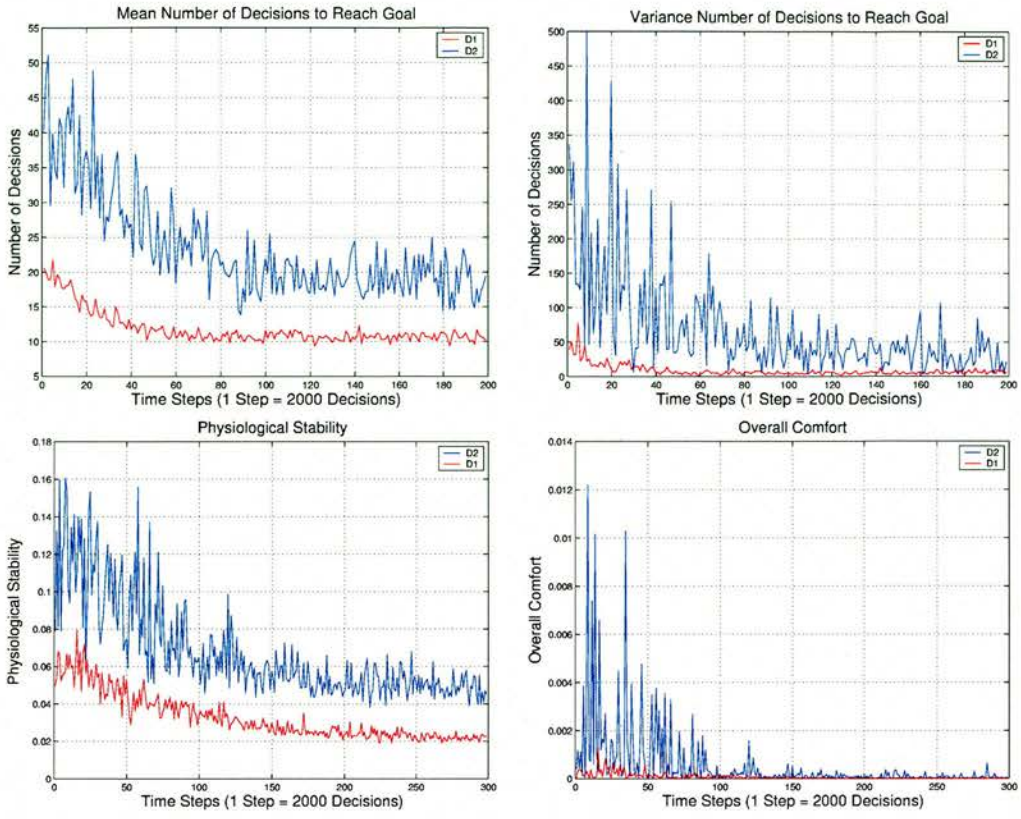


Figure 5.5: Top Graphs: Evolution of the Number of Decisions to Reach Goal for two distributions of affordances, D1 and D2. The graphs represent the mean and variance, left and right, respectively. Bottom Graphs: Evolution of the Physiological Stability and Overall Comfort (viability indicators), left and right, respectively.

overall comfort improve with the learning process, until a stationary value is reached. For both the metrics, learning is demonstrated to make the system more stable and more efficient in terms of its internal physiology. Final physiological values around 0.5 and 0.02 have been reached for case D1 and D2, respectively. Overall comfort values reached are under 10^{-3}. These results demonstrate that the actor-critic has been able to learn appropriate policies, and that making the physiology stability is very much related to the consecution of a short behavioural pattern.

Behaviour Assessment The correspondence between drives and behaviours is shown in figure 5.6. The learning process is presented in terms of *learning cycles*, shown in the left and central graphs. The learning is divided into episodes, starting with the reset of the homeostatic variables at random values and ending when these reach the optimal zone. Graphs on the left and in the centre illustrate the initial and final cycles of the drives and of the policy functions

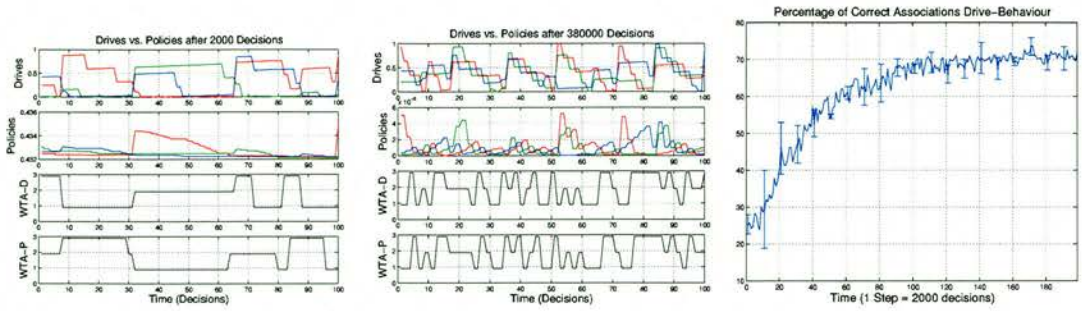


Figure 5.6: These graphs characterise the evolution of the relationship between the drives and the behaviours throughout the simulation for the case of an environment with uniformly distributed affordances (D1). The graphs, for the left and central case, show, from top to bottom: the level of each drive, the level of each associated policy value, the winning Drive, the winning policy (each drive and its related behaviour is drawn with the same colour (Red for Hunger-Eat (behaviour 1), Green for Tiredness-Rest (behaviour 2), Blue for Restlessness-Touch (behaviour 3)). The graph on the right shows the percentage of agreement between the winner drive (WTA-D) and the winning policy function (WTA-P) along the simulation.

(expressing preference for one or another behaviour. For the given situation, each drive can be satisfied by a single behaviour). The right match has been stressed by using the same colour to depict the drive and its related behaviour (red for the Hunger/Eat, Green for Tiredness/Rest and Blue for Restlessness/Touch, drives/behaviours, respectively). The learning cycles are shown at the beginning and end of the simulation, left and central graphs, respectively. The procedure for selection consists of selecting the behaviour whose policy exhibits the highest value. Keeping this in mind, the actor-critic should learn that the most urgent drive has to be served first, hence if a behaviour exhibits the highest intensity (largest related motivation), its related motivation should also be the most intense. The two bottom graphs in the left and central groups show the winner drive (WTA-D) and of the policies (WTA-P), respectively. This is coherent with the results: the graphs on the left (beginning of the simulation) show a high level of discordance between both WTA graphs, but this decreases gradually during the simulation, cf. central graph on figure 5.6 (end of the simulation). The graph on the right hand side on figure 5.6 shows an average over 20 simulations of the percentage of agreement between the behaviour related to the winner drive (the one that should be executed according to the aforementioned criterion of selection) and the behaviour selected in simulation. The evolution of this matching starts from 20%, and reaches a final value ca. 70%. Higher values are not likely to be reached due to the 20% of random selection and due to the numerical accuracy boundaries of the neural networks.

The actor-critic has demonstrated, in this context, its ability to *relate drives and behaviours in such a manner that reward is maximised.* Hence, to define contributions towards stability as

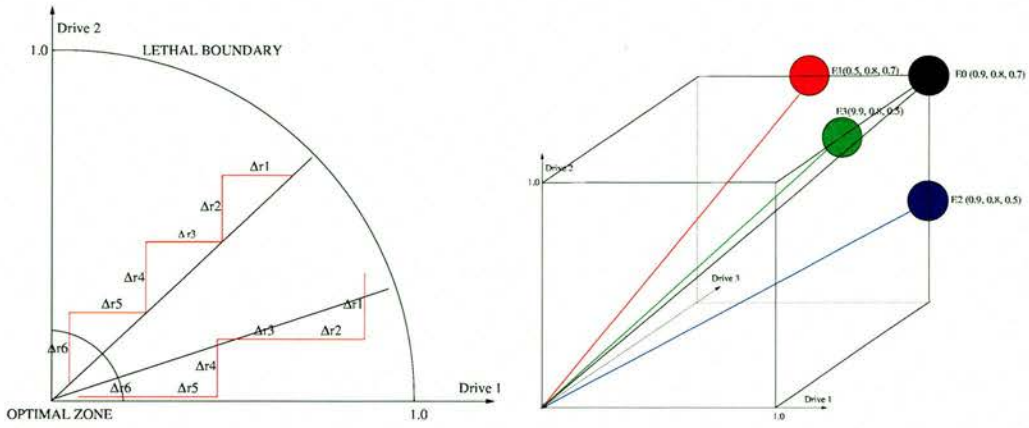


Figure 5.7: The picture on the left is an illustration of an example of two sets of effects (in red) due to behaviour execution. These start at a random point in the agent's physiological space and end in its optimal physiological zone. Both the effects are due to the execution of a sequence of behaviours. The picture on the right shows the four initial states considered for experiments shown in graphs 5.8 and 5.9.

positive suffices to come up with a strategy for selection that relates every need of the agent to an appropriate behavioural response. This suggests that the actor-critic algorithm is able not only to learn to relate stimulus to responses, but also capable of learning to select the most appropriate behaviour that mediates the maximisation of reward. I need to stress that these results are so far solely valid for the definition of reward that interprets as good those actions leading to an overall more stable physiological situation. This may also suggest that, in principle, the thesis supported by Schultz (1998) arguing that some parts of the brain behave as an actor-critic to mediate Pavlovian learning can also be extended to the instrumental case, and that this may be one of the mechanisms within the strategy for selecting the behaviour to execute next in biological beings. In addition to this, this is also consistent with the view of Rolls (2003) and Sutton and Barto (1998) on behaviour selection: at each level of the hierarchy, the behaviour to execute next tends to be within a sequence leading to the highest cumulative reward. In this particular case, this is equivalent to selecting the behaviour whose associated drive expresses the highest urge for compensation 70% of the time. Hence, the most urgent drive is served first.

To further test the consistency of the aforementioned results, I have performed four different simulations with the same parameters used in these last experiments. However, a learning and an assessment phase alternate every 2000 decisions, during which the cycle starts from a fixed position in the agent's physiological space (cf. figure 5.7). Four different initial physiological values have been deliberately chosen to test the behavioural response for four different physiological imbalances. Each assessment phase extends over 1000 decisions. During ev-

ery assessment phase, values of effectiveness and cycles of execution of behaviour have been recorded and plotted in figures 5.8 and 5.9.

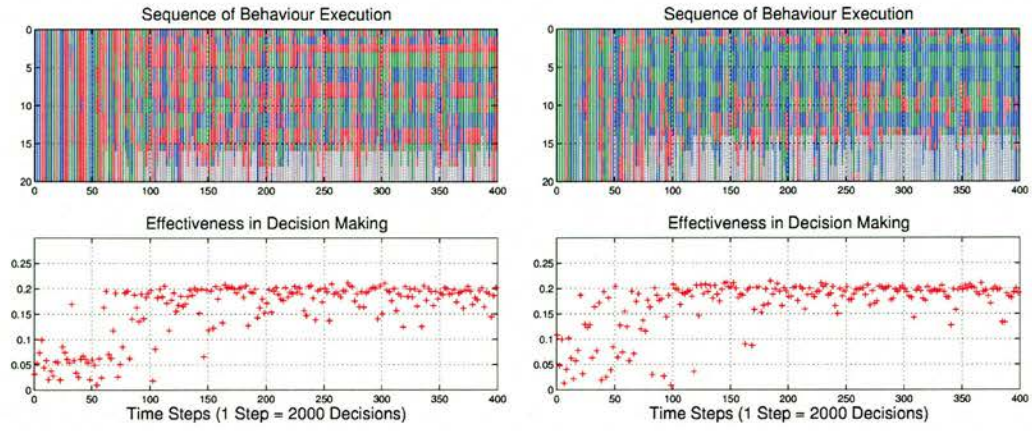


Figure 5.8: Representation of the Evolution of the Behavioural Cycles throughout the Learning Process. Two different initial physiological states have been chosen. For each case, the top graph shows the evolution of the sequence of behaviours executed from the same initial physiological state to reach the optimal zone in a 2-D physical space. The sequences of behaviour execution start on top of the figure, and flow down the y-axis. The colours are Red for Grasp, Green for Shelter, Blue for Touch. Grey means that the goal has been reached. The bottom graph of each pair shows the evolution of the effectiveness metric, calculated according to equation 5.8. Initial states (hunger,tiredness,restlessness) are: (0.9, 0.8, 0.7) and (0.5, 0.8, 0.9), left and right, respectively.

These results are necessary to reach an *element of ethological comparison*, inspired by McFarland's physiological cycles (McFarland and Spier, 1997). So far, cycles of execution have been mainly portrayed in two dimensions only. This representation can be used in N dimensions. Furthermore, it is accompanied by a metric of effectiveness of behaviour execution designed, *ad hoc*. This is introduced by equation 5.8 (Δr_i being the size of the effect of the execution of the i^{th} behaviour within the cycle and N the total number of executions during that cycle) and figure 5.7. Effectiveness is defined as a quotient between the sum of the effects from the beginning of the cycle until the optimal zone is reached, normalised by the number of executions entailed in the process. Hence, if most executions have an effect, the value of the metric will approach sequentially Δ (0.3 in this case).

$$Effectiveness = \frac{1}{N}\sum_{i=0}^{N-1}\Delta r_i \tag{5.8}$$

Hence, if the agent is learning the shortest path towards the optimal zone, the sequence should resemble a straight line from the initial state to the origin of the physiological state.

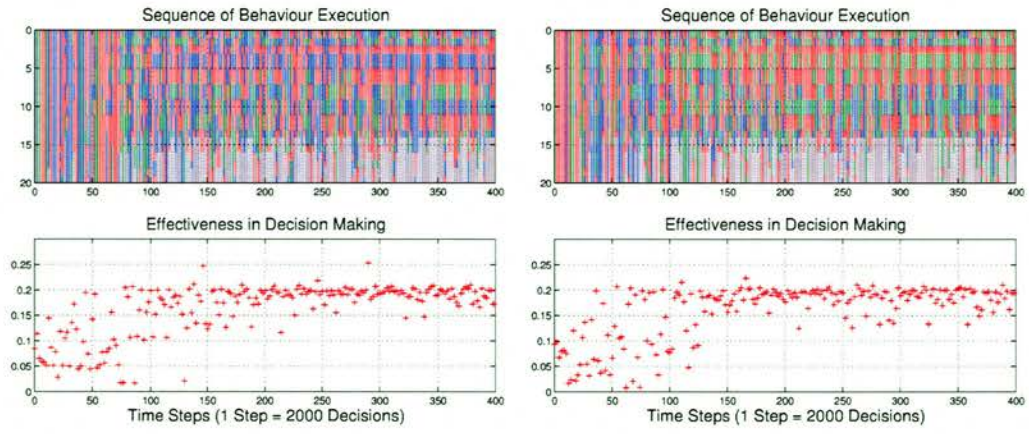


Figure 5.9: Representation of the Evolution of the Behavioural Cycles throughout the Learning Process. Two different initial physiological states have been chosen. For each case, the top graph shows the evolution of the sequence of behaviours executed from the same initial physiological state to reach the optimal zone. The sequences of behaviour execution start on top of the figure, and flow along the y-axis. The colours are Red for Grasp, Green for Shelter, Blue for Touch. Grey means that the goal has been reached. The bottom graph of each pair shows the evolution of the effectiveness metric, calculated according to equation 5.8. Initial states (hunger,tiredness, restlessness) are: (0.9, 0.5, 0.8) and (0.9, 0.8, 0.5), left and right, respectively.

The effectiveness metric has been designed to be independent of the initial state. In fact, only the proportion between effective and non-effective behaviours executed will matter, e.g., a sequence entailing 5 effective and 2 non-effective behaviours will have the same effectiveness as a sequence entailing 10 effective and 4 non-effective behaviours ($\frac{5+2}{7} = \frac{10+4}{14}$). Furthermore, the sequence of behaviours executed in the second case (arbitrarily chosen) at each evaluation phase has been stored.

Both metrics, the behavioural sequence and the effectiveness have been plotted in graphs 5.8 and 5.9. The first (figure 5.8, left graph) aims at offering a complete behavioural sequence starting always from a physiological state where deficits are very high (0.9, 0.8, 0.7). Behaviours are identified by colour. The top figures show that the sequence of behaviours at the beginning of the simulation exhibits very long cycles, since a sensible pattern has still not been learnt and the same behaviour is repeated aimlessly. However, at around 100 (X-axis), cycles start shortening, and a pattern of execution of behaviours starts to emerge (with some variations from value 250 onwards). Red, green and blue behaviours can be seen repeating in the y-direction, indicating that all three behaviours are chosen in turn to decrease the deficits in each of the three drives. This also relates to the value of effectiveness displayed underneath, which increases from the same x-value onwards.

Similarly, the three other cases of study aim at showing the sequence of behaviours when the starting physiological state is most deficitary in each of its 3 dimensions. Consistently, the initial states are, for Hunger, Tiredness and Restlessness, respectively, (0.5, 0.8, 0.9), (0.9, 0.5, 0.8) and (0.9, 0.8, 0.5). Hence, for each case, it can be observed that most of the behaviours in the behaviour sequence at the end of the simulation correspond to those behaviours whose deficit is the highest. Hence, for the first case (figure 5.8, right graph), most behaviours are touch and shelter (blue and green), for the second case (figure 5.9, left graph) they are grasp and touch (red and blue) and for the third case (figure 5.9, right graph) they are grasp and shelter (red and green). The stable value of effectiveness corresponds to the moment when the pattern of effective behaviour execution arises.

The metrics introduced in this section, to assess *learning*, *behaviour selection*, *physiological stability* and *cycles of execution* will be used in a set of ethologically relevant cases introduced next.

5.4 Learning to Select Consummatory Behaviours

This section is devoted to reporting studies on the effect of the *distribution and availability* of resources in the learning process. The *hypothesis* is that both the environment and the internal drives modulate the definition and the learning of the motivational state, wherein decisions are made. This is, the distribution of resources (in the form of affordances) should control the range of behaviours that the actor-critic learns, ranging between the stimulus and the motivation driven.

Furthermore, the internal drives will also bias the selection in one or another manner, e.g., when hunger is the dominant drive, the agent should strongly bias the selection of behaviours compensating that need. However, this can only happen if the object faced affords to perform those behaviours. In this respect, I introduce two particular cases of distribution: a stimulus driven environment (distribution K1, cf. figure 5.10), where every object affords a single behaviour, and a motivation driven distribution where every object affords every behaviour (distribution K2).

In a *stimulus driven environment* the agent has the choice of executing the behaviour afforded by the object nearby to compensate its internal drives. The selection of other behaviours will report a loss of reward to the agent. Therefore, since the actor-critic aims at accumulating reward, its final policies for selecting behaviour should demonstrate that the external stimuli dominate over the internal. For our experiments, a stimulus driven environment has been designed by distributing the availability of the resources in the environment to every object as a function of their size in a non-overlapping manner, cf. figure 5.10.

The converse case to this environment is the *motivation driven environment*. An agent in

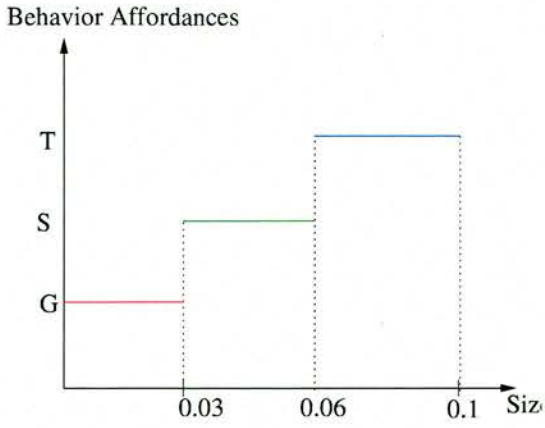


Figure 5.10: K1 Distribution of affordances. The line indicates the interval of object sizes where that affordance is 1.0. G, S and T stand for grasping, shelter and touch, respectively. Object sizes range from 0.0 to 0.1.

this second type of environment should not make decisions conditioned on external stimuli, but on the internal drives only. I hypothesise, within the context of the actor-critic, that both stimulus and motivation driven patterns of behaviour selection should emerge naturally via interaction with the environment. The argument supporting this is that the actor-critic maximises the expectation of reward, and this should be at its maximum when every decision exerts an effect towards the stability of the internal physiology. This occurs when the behaviour has a compensatory internal effect on the internal drives of the agent, therefore when the behaviour selected matches one of the affordances offered by the object nearby.

In this light, convergence can be studied in terms of the number of decisions required to lead the physiological state to the viability zone, which decreases over time (cf. figure 5.11 mean and variance, top-left and top-right, respectively). Furthermore, I have also measured the evolution of the physiological stability and overall comfort throughout the simulation, which decreases while the convergence increases, hence demonstrating that convergence and stability of the behaviour selection pattern are learnt by the actor-critic.

The top pictures of figure 5.12 show a sample of the cycles at the beginning and end of the simulation, left and centre, respectively. It can be observed that once the policies for selecting behaviours have been learnt, the perceived affordances match the behaviours selected by these policies. The set of graphs in the third and fourth rows of each set compare the learning performance by showing the affordances offered by the objects aside to the policy functions to select behaviours at the beginning and end of the simulation, left and middle graphs, respectively. The graph on the right hand side shows the evolution of the percentage of matching between the winner policy and the winner drive (labelled as motivation driven) for the environments K1 and K2, top and bottom, respectively. Furthermore, for the K1 environment only, the evolution of the matching between the winner policy and the affordance offered at every encounter

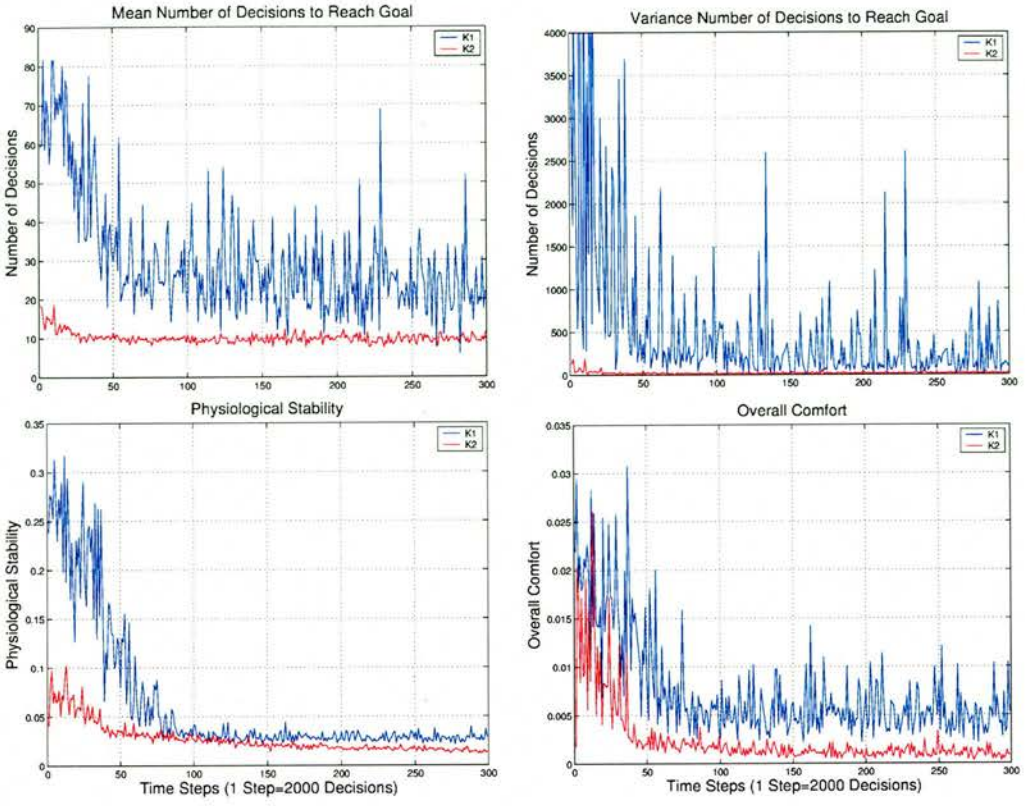


Figure 5.11: Top Graphs: Evolution of the Number of Decisions to Reach Goal for distributions K1 and K2, see figure 5.10. The graphs represent the mean and variance, left and right, respectively. Bottom Graphs: Evolution of the Physiological Stability and the Overall Comfort (viability indicators), left and right, respectively.

throughout the simulation is also shown (labelled as incentive driven).

The results show that learning is more difficult in the environment K1 than in the environment K2. The explanation for this is that the only behaviour afforded by each object in the environment K1 is the one afforded by the environment, hence there is no choice if reward is to be maximised. The learning process is aimed at finding the right distribution such that reward is maximised. The only way is to effectively profit from every opportunity, therefore to disregard the internal needs and to choose to execute the behaviour offered by the object nearby at every encounter.

The level of agreement between the curves on the right hand side of figure 5.12 shows that the actor-critic becomes stimulus driven after 12×10^4 decisions. This number of decisions may seem large if compared to biological data (Schultz et al., 1997). However, it is important to consider that its biological counterpart has been hypothesised to incorporate a model from the pre-frontal cortex, whereas our learning is solely based on TD-Learning (which is model-

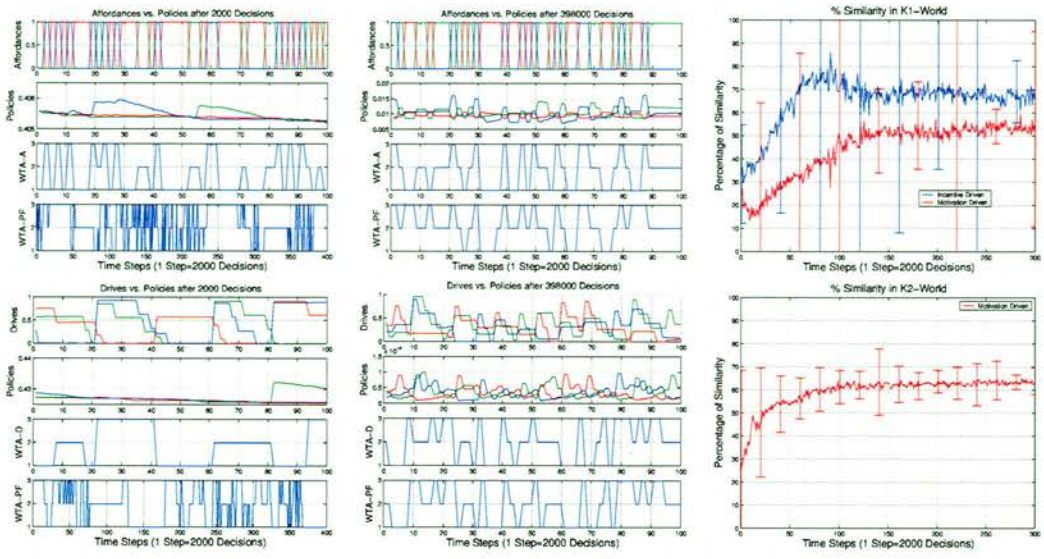


Figure 5.12: These graphs characterise the evolution of the relationship between the drives and the behaviours throughout the simulation for the case of an environment with uniformly distributed affordances (K1), cf. figure 5.10. The graphs, for the left and middle case, show, from top to bottom: the level of each drive, the level of each associated policy value, the WTA of the Affordances (i.e., the affordance available at the moment of decision making), the WTA of the Policies (each drive and its related behaviour are drawn with the same colour (Red for Hunger-Eat, Green for Tiredness-Rest, Blue for Restlessness-Touch)). The graph on the right hand side shows the percentage of coincidence between the WTA-A (the affordance) and the behaviour selected (WTA-P). The top graph compares distributions K1 and K2. The lower case only refers to distribution K2.

free). Furthermore, the results show that for environment K1 (incentive environment), the behavioural pattern becomes stimulus driven for around 60-65% of the encounters after 240×10^4 decisions. To this extent, the agent is exhibiting a reactive behaviour. The reasons for this are that 20% of selection is at random for exploratory purposes and that the numerical accuracy of the neural networks estimating the motivational state is limited. Furthermore, the decay of the homeostatic variables has been fixed for these experiments at ($\tau = 10^{-3}$), which provokes a noticeable effect of *satiation* when the policies are learnt. If the behaviour afforded by the object encountered corresponds to a drive already satisfied, the effect of its execution is close to 0. The satiation effect contributes to reduce the metric by a significant 10-15%.

Graphs in figure 5.13 and 5.14 show the cycles for the case of an incentive environment K1. In this case, it is necessary to consider that every object in this environment offers a single affordance and that interaction occurs on a random basis. Hence, despite the learnt

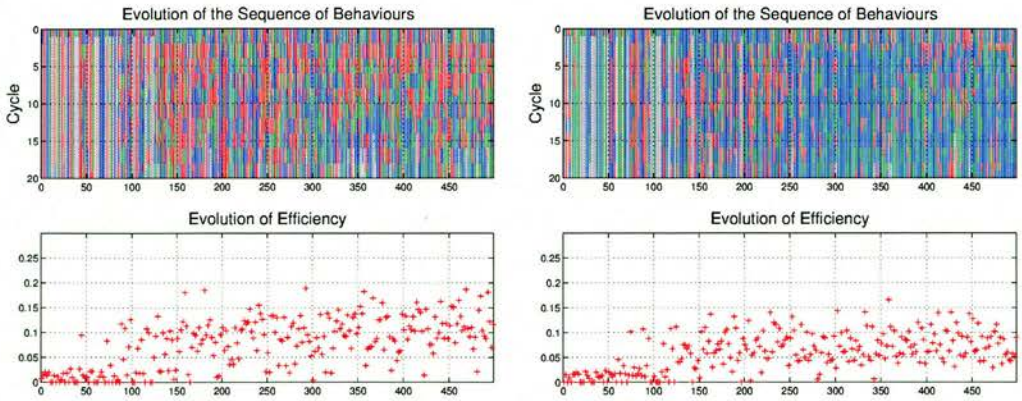


Figure 5.13: Representation of the Evolution of the Behavioural Cycles throughout the Learning Process for the case of reactive environment. Two different initial physiological states have been chosen. For each case, the top graph shows the evolution of the sequence of behaviours executed from the same initial physiological state to reach the optimal zone in a 2-D physical space. The sequences of behaviour execution start on top of the figure, and flow down the y-axis. Only the first 20 behaviours composing the cycle are displayed (it is expected that the cycle will be shorter than that by the end of the simulation). The colours are Red for Grasp, Green for Shelter, Blue for Touch. Grey means that the goal has been reached. The bottom graph of each pair shows the evolution of the effectiveness metric, calculated according to equation 5.8. Initial states (hunger,tiredness,restlessness) are: (0.9, 0.8, 0.7) and (0.5, 0.8, 0.9), left and right, respectively.

patterns being shorter and rewarding to the agent, these should be different among themselves. Accordingly, these figures show that the behavioural patterns shorten throughout time due to the learning. However, the final behavioural cycles vary substantially among themselves in every one of the four situations considered (the colours among behavioural cycles of the same figure alternate without any certain regularity). This suggests that the patterns respond to the behaviour affordances offered at every interaction, and therefore that the behavioural responses are reactive.

5.4.1 Behaviours with Double Effect

The results in the previous section have demonstrated that the learning architecture does learn a stable behavioural pattern for environments endowed with distributions of affordances ranging from the stimulus driven to the motivation driven. Furthermore, the behavioural patterns have demonstrated that they suffice to maintain the stability of the internal resources for the given conditions. However, the reach of these conclusions is bounded by the use of a unique correspondence between every drive of the agent and a particular behaviour.

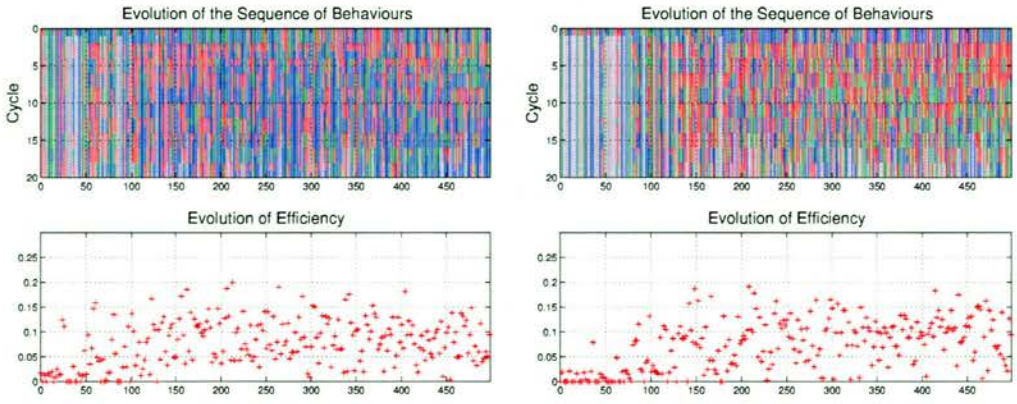


Figure 5.14: Representation of the Evolution of the Behavioural Cycles throughout the Learning Process. Two different initial physiological states have been chosen. For each case, the top graph shows the evolution of the sequence of behaviours executed from the same initial physiological state to reach the optimal zone in a 2-D physical space. The sequences of behaviour execution start on top of the figure, and flow down the y-axis. The colours are Red for Grasp, Green for Shelter, Blue for Touch. Grey means that the goal has been reached. The bottom graph of each pair shows the evolution of the effectiveness metric, calculated according to equation 5.8. Initial states (hunger, tiredness and restlessness) are: (0.9, 0.5, 0.8) and (0.9, 0.8, 0.5), left and right, respectively.

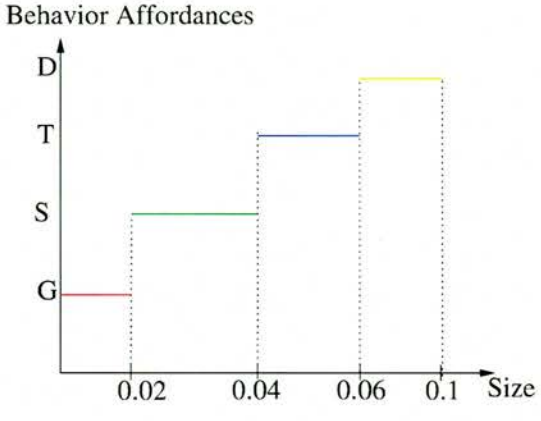


Figure 5.15: L1 Distribution of affordances. The line indicates the interval of object sizes where that affordance is 1.0. G, S, T, D stand for grasping, shelter, touch and drink, respectively. Object sizes range from 0.0 to 0.1. Distribution L2 is a uniform distribution, where every behaviour is afforded by every object.

In order to extend these conclusions, this section adds asymmetry to the relationship between drives and behaviours by adding a new behaviour to the agent's repertoire. Furthermore, this new behaviour has been chosen in such a manner that when executed, it concurrently diminishes two deficits (hunger and tiredness); therefore introducing multiple solutions for the

compensation of the agent's physiological deficits. Now it will be possible for the agent to either sequentially execute to rest and to eat or to execute the new behaviour to reduce the aforementioned drives.

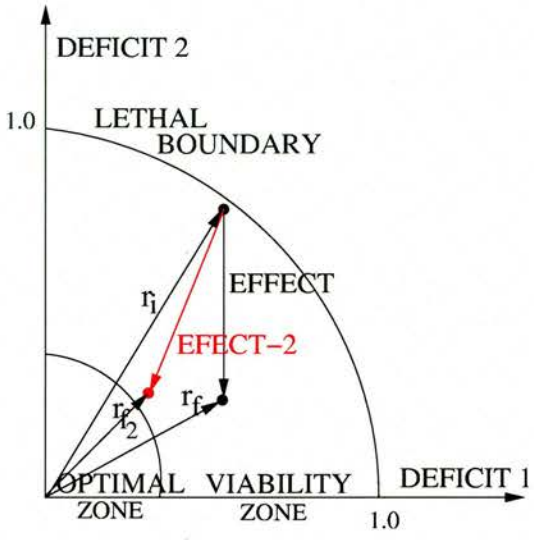


Figure 5.16: This figure shows the physiological effect of executing a behaviour that compensates two drives (in red) in comparison with the effect of a behaviour that compensates only one drive (in black). The cartesian representation stands for a two dimensional physiological state, where x and y are the two deficits. Optimally, the agent's deficits should remain within the optimal zone.

The top graphs in figure 5.17 show the evolution of the mean length of the behavioural cycle throughout the simulation. The cycle starts at a random value of the physiological state, and ends when the agent's drives are close to zero (have reached the optimal zone). The two lines correspond to the two distributions of affordances L1 and L2 described above (cf. figure 5.15), in red and blue, respectively. In a complementary fashion, the bottom graphs of the same figure show the evolution of the viability indicators, physiological stability and overall comfort. Both the length of the behavioural cycle and the viability indicators exhibit a similar tendency that responds to a similar pattern to those shown in previous sections. Once the actor-critic has learnt a rewarding policy, the length of the cycle diminishes, and the viability indicators reduce their values, meaning that the agent has learnt to compensate its needs by selecting the appropriate behaviour within the agent's repertoire. The graphs are the result of averaging these metrics over 20 simulations.

The influence of adding one more behaviour has some effect, since the mean length of the behavioural cycle is shorter than for the case of having only three behaviours (you can compare with graphs in figure 5.11). Likewise, the viability indicators exhibit smaller values, meaning that the addition of the new behaviour has increased the ways to correct a deficit and

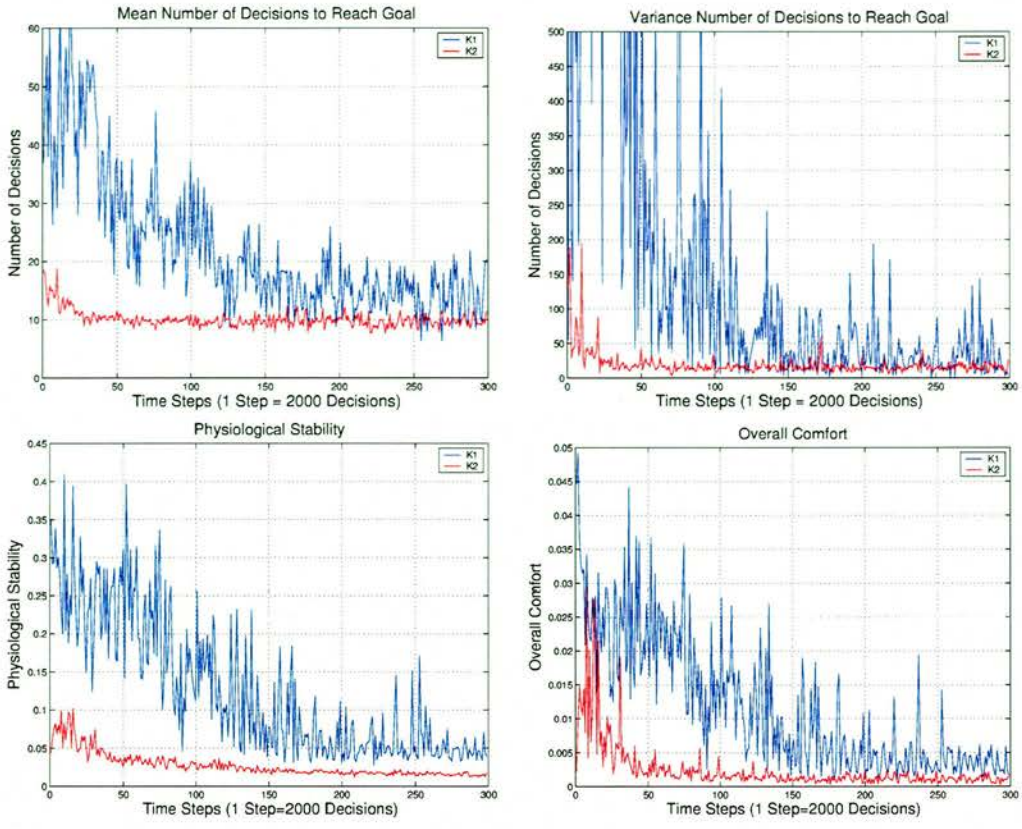


Figure 5.17: Top Graphs: Evolution of the Number of Decisions to Reach Goal for the distribution of affordances L1 and L2. The graphs represent the mean and variance, left and right, respectively. Bottom Graphs: Evolution of the Physiological Stability and the Overall Comfort (viability indicators), left and right, respectively.

consequently lowered the final viability values. However, in addition to these beneficial effects, adding one more behaviour has also had an additional cost in terms of computation, since now the actor-critic's state space extends over 4 dimensions, which explains the rise of the necessary number of decisions to reach convergence, from 24×10^4 to 30×10^4.

Furthermore, given the lack of symmetry of the responses of behaviours to the homeostatic variables, the *analysis of cycles* for this case must differ from the previous cases. As for the previous experiments, the trials shown in this section aim at analysing the policies learnt by the actor-critic. However, in order to attain this goal for this case, the comparison to motivation-driven or stimulus-driven policies is simply not possible when every situation offers more than one possible behavioural response. To solve this problem I need to relate to the formulation of principles during the literature review. When introducing the actor-critic (see section 2.5.3), I stated that the combination of stimuli to select behaviours would respond neither to a multiplicative nor to an additive formulation of stimuli. Unlike this, the selection of behaviours in

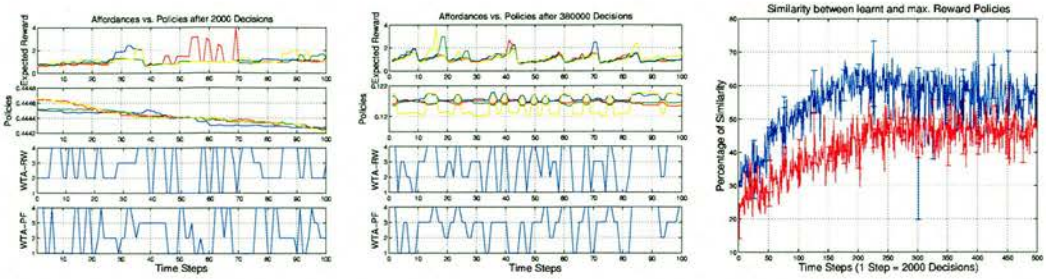


Figure 5.18: These graphs characterise the evolution of the relationship between the theoretically most rewarding decision and the decision that the agent has learnt throughout the simulation. The graphs,for the left and central case, show, from top to bottom: the level of theoretical functions (driven by reward), the actor-critic policies, the WTA of the theoretical policies, the WTA of the real policies (every behaviour are drawn with the same colour as its policy (Red for Eat, Green for Rest, Blue for Touch, Yellow for Drink, Cyan for Avoid). The graph on the right shows the percentage of agreement between the theoretical and the stimulus driven throughout the whole simulation.

our model exhibits patterns arisen via interaction with the environment. This is in turn assessed by the reward or punishment resulting from the execution of its behaviours, related to the stability of the agent's internal milieu. Therefore, if the best patterns are the most rewarding, it also makes sense to compare this theoretical assumption to the real patterns obtained in simulation. Therefore, our analysis has consisted of calculating the theoretically best behavioural response at every encounter in terms of reward (top graphs, centre and middle groups, beginning and end of the simulation, respectively) at every encounter and of comparing this with the decision that the agent really made. The bottom graphs for the same groups show the theoretical winner in terms of reward and the real responses learnt by the actor-critic. By comparing them, I can observe their resemblance at the end of the simulation (central group of graphs).Therefore, this demonstrates that if the object encountered affords two different behaviours to be performed, the system executes the behaviour providing the most reward. For this case, the behavioural response may also differ if the level of one homeostatic variable or another is already sated. In equal affordance conditions, the actor-critic should learn to attend first to the homeostatic variable exhibiting the highest deficit. These ideas reflect on the graphs on the right hand side, which compare the degree of similarity of the theoretical and real decisional pattern for the distributions L1 and L2 throughout the simulation. The satiation of some homeostatic variables (therefore the fast pace of decision making of the agent compared to the internal consumption) leads to results that demonstrate that making the most rewarding decision is not necessary approximately 40% of the time (cf. blue graph). The same effect can also be appreciated for the red graph (L2 distribution).

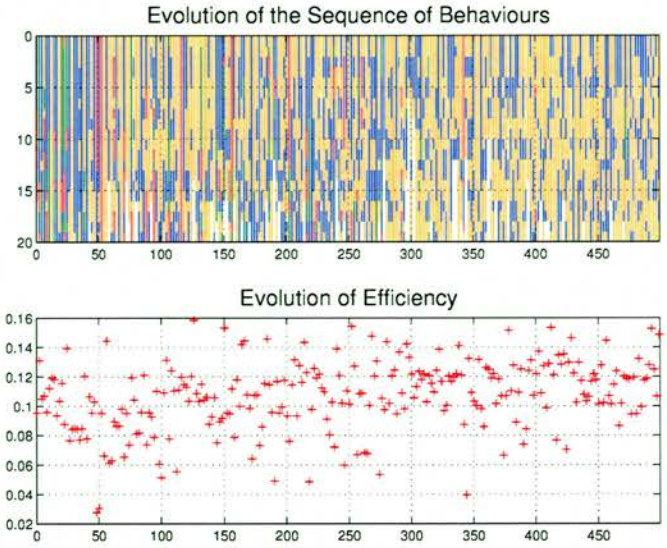


Figure 5.19: Representation of the Evolution of the Behavioural Cycles throughout the Learning Process. Four different initial physiological states have been chosen. The bottom graph of each pair shows the evolution of the effectiveness metric, calculated according to equation 5.8. The initial state (hunger,tiredness,restlessness) is (0.9, 0.5, 0.8). The colours stand for: Red for Eat, Green for Rest, Blue for Touch, Orange for Drink.).

The *analysis of cycles* has been performed for a single case. The orange areas in figure 5.19 demonstrate that the actor-critic has been able to incorporate a fourth behaviour (labelled as "to drink"), which compensates two different homeostatic variables. The final length of the cycle is around 7. Furthermore, it can be observed that the final pattern consists of two behaviours, to drink (orange colour), which compensates the variables nutrition and stamina and the blue behaviour (touch), which compensates curiosity. This also indicates that this is the most rewarding path to reach the viability zone and at the same time, the shortest path to this end.

Beyond the learning effect, it can also be observed that the length of the simulation has significantly increased due to the addition of one more behaviour. This is due to the extension, by one more dimension, of the search space. As for the former cases, this could be palliated by using an internal model in the need of scaling to higher dimensional spaces.

5.5 Learning Policies in an Asymmetric Architecture with Appetitive and Consummatory Behaviours

Previous experiments in this chapter have addressed the learning of patterns of decisions among consummatory behaviours only. However, the integration of these with appetitive behaviours

in the competition of the agent's actuators is still an unclear matter.

This section does not intend to provide a complete answer to this question, the reach of which is very extensive. However, it does intend to highlight that the actor-critic can, under some circumstances, learn to integrate appetitive and consummatory behaviours in a successful manner. I argue that there is a motivation related to each behaviour, both appetitive and consummatory, within the agent's repertoire and that it is the beneficial combination in terms of reward of appetitive and consummatory behaviours which makes this combination possible.

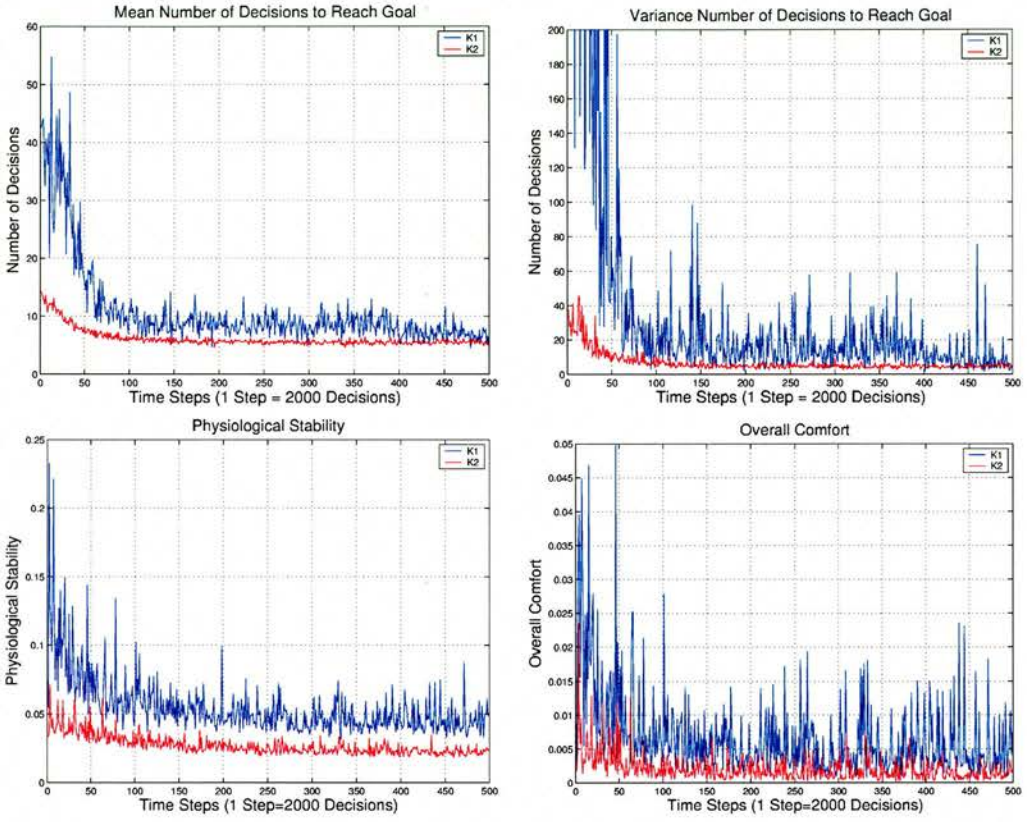


Figure 5.20: Top Graphs: Evolution of the Number of Decisions to Reach Goal for the distribution of affordances presented by figure 5.10 with an agent endowed with an appetitive behaviour: to avoid. The graphs represent the mean and variance, left and right, respectively. Bottom Graphs: Evolution of the Physiological Stability and the Overall Comfort (viability indicators), left and right, respectively.

5.5.1 Integration Appetitive and Consummatory Behaviours

The integration of appetitive and consummatory behaviours is a problem extending in several directions (Toates and Jensen, 1990). Although this problem has been addressed during the last decade by several authors (Blumberg, 1997, 1994; Tyrrell, 1993), an appropriate solution

to the problem is still missing. The main difficulty is the lack of perspective when formulating the integration of appetitive and consummatory behaviours in a single motivational framework, where behaviours compete with one another to gain control of the actuators. Related to this, the first difficulty is the disagreement on the need of integrating both sorts of behaviours in the same motivation driven framework, since some authors suggest that appetitive behaviours are not motivation driven. In this respect, I adhere to the view of McFarland and Spier (1997) arguing that "all behaviour of any animal will by definition will be guided by a motivational state". Furthermore, I suggest that the integration of both sorts of behaviours is possible if considered in a reinforcement learning framework. In such a framework, the execution of behaviours can be assimilated to the transition between physiological states, for whose execution there no need of immediate reward.

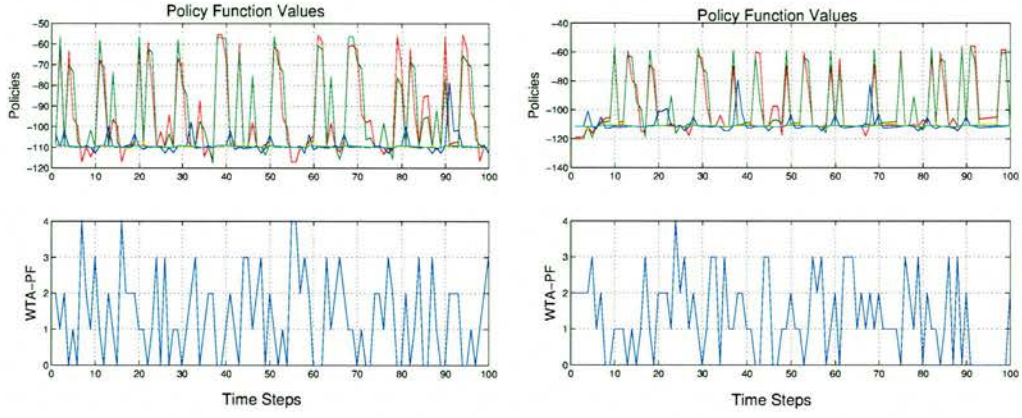


Figure 5.21: Top Graphs: Policy functions at the end of two different simulations, left and right, respectively. The colours stand for the behaviours grasp (red), rest (green), touch (blue), drink (cyan), avoid (orange). Every behaviour has also been labelled by numbers, from 0 to 4, assigned in the same order, respectively. The bottom graphs display the WTA of the preference functions shown in the figures above at the end of the same two simulations.

In order to test this, a fifth behaviour (to avoid) has been included in the architecture. Unlike its competitors, it does not provoke any compensation of the physiological state. However, in the same way as the rest of the behaviours, its execution does encompass a consumption of internal resources. It is argued that this assumption is reasonable, since even though its execution takes a shorter time than any consummatory behaviour, and does not imply anything but the avoidance of the object, it still needs resources. Hence, this behaviour is competing in equal conditions to consummatory behaviours at the moment of decision making though it does not affect the internal physiology in the same manner. While the consummatory behaviour may exert an internal deficit compensation, the appetitive behaviour is only helpful to situate the agent to appropriately select a consummatory behaviour.

The *experimental framework* has consisted of the same environment L1 used in the previous section. However, in addition to the affordance distribution of behaviour affordances, every object affords to be avoided. The same agent, endowed with three homeostatic variables, nutrition, stamina and curiosity, and three drives, hunger, tiredness and restlessness has been used for simulation. The effect of every consummatory behaviour (to grasp, to shelter, to touch, to drink) is 0.3 (α_{ik}) in equation 4.1, and their decay constant of every homoestatic variable is $\tau = 10^{-3}$.

Figure 5.21 shows the policy functions at the end of two different simulations and demonstrates that the behaviour avoid has been learnt is selected in some situations. The avoid policy values (orange line) has the same mean value as the resting of the consummatory behaviours. This common mean value indicates that these are all competing in equal conditions and that each of them is, depending on the input state, deliberately selected. In fact, it can be observed that the behaviour four (to avoid) is selected on several occasions. In order to analyse the circumstances of this decision making, the final cycles of this experiment are shown in figure 5.22. The patterns displayed in this figure correspond to two trials using the same simulation parameters. The top patterns show an overall view throughout both simulations, from beginning to end. The bottom patterns show the final cycles of this simulation. The colours in this figure are assigned in the following manner: red (grasp), green (shelter), blue (touch), magenta (drink) and yellow (avoid). In both bottom patterns, it can be observed that the behaviour avoid has been integrated into the decision patterns, although it is not as frequent as its consummatory competitors. This demonstrates that this behaviour can be convenient from the perspective of reward. In other words, for a certain distribution of affordances, there will be situations where the need of the agent cannot be met by the object encountered. In this moment, instead of attempting the execution of any consummatory behaviour, which would fail and leave the need unsatisfied, it may be more intelligent to avoid the object and to search for another object, more appropriate to the agent's needs. This decision will spare the consumption of internal resources. The results show that this reasoning emerges naturally from the relationship to the environment in the agent's behavioural patterns.

The results introduced in this section do not justify every case of appetitive behaviour; whose integration with consummatory ones should be analysed on an individual basis. However, it has demonstrated that the actor-critic, via the principle of ecology, can learn to integrate this single appetitive behaviour in equal conditions to its consummatory competitors in its decision patterns. This issue completes the cases of study considered for the actor-critic and confirms its adaptivity to a variety of architectures and environments. This issue is further completed in the chapter's conclusion.

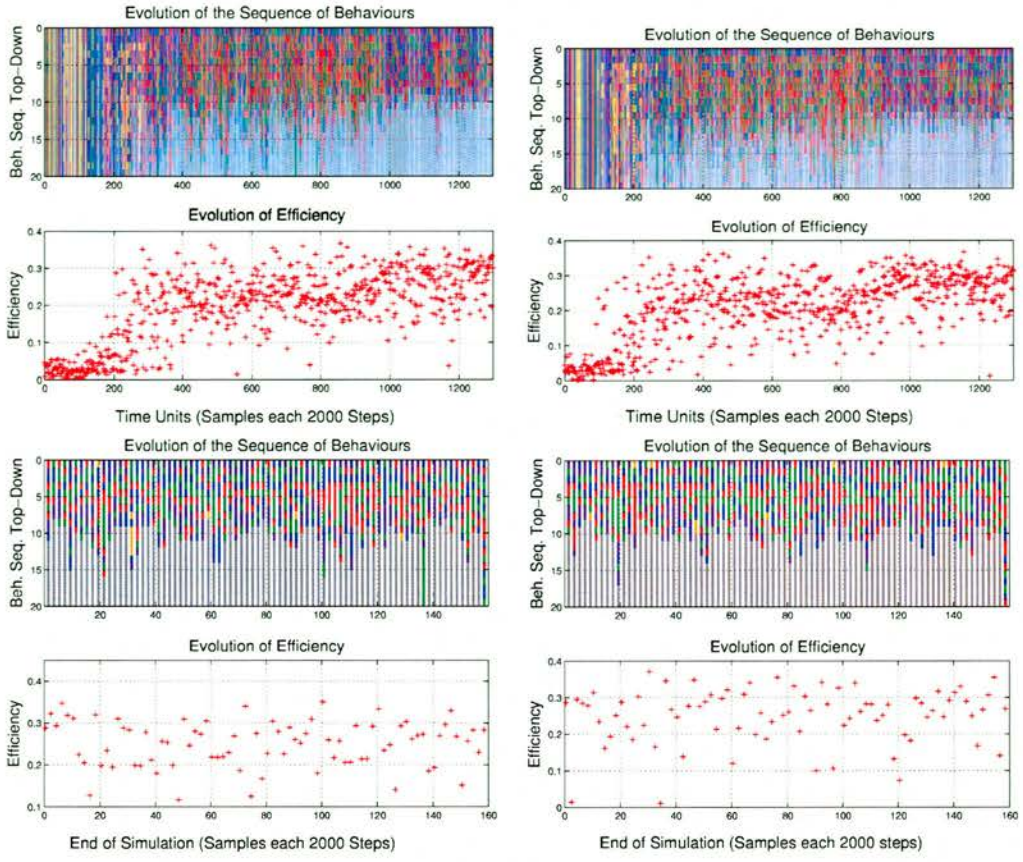


Figure 5.22: Representation of the Evolution of the Behavioural Cycles throughout the Learning Process. The learning cycle has always been started at the same initial physiological set of values: (hunger,tiredness,restlessness) are (0.9, 0.5, 0.8). The top graphs show the evolution of the execution cycle throughout the whole simulation :every cycle starts at a random value of the agent's physiological state and finishes when the optimal zone has been reached. The colours stand for the behaviours grasp (red), shelter (green), touch (blue), drink (cyan), avoid (orange). The cycles below show the last 150 behavioural cycles of the cases shown above. The dotted graphs in red show the evolution of the effectiveness of the behaviour executions, calculated according to equation 5.8.

5.6 Conclusion

This chapter has introduced a biologically inspired model that is designed to learn to select behaviours. I have addressed high-level behaviour selection by implementing the hypothesis of DA mediating Pavlovian learning in the Basal Ganglia (BG) and extending this to the case of instrumental learning. In order to test this hypothesis, a learning framework including the actor-critic has been modelled together with a series of modules simulating an artificial physiology, which have been integrated in a single architecture. The architecture, implemented in

a simulated robot, has been tested in a set of simulated environments reproducing a variety of ethologically significant situations. The environments have been characterised according to their distribution and availability of resources (affordances); these range from the abundant to the scarce and from the stimulus driven to the motivation driven.

The learning architecture has been conceived as an extension of the perception architecture proposed in the previous chapter, however focusing on the behaviour selection and related processes that modulate to adapt to the environment. To this end, the architecture has not imposed any further constraint on the combination of stimuli than the relationship to the environment that is perceived via the reward that results from executing behaviours. Therefore, in this case there is no specific formula to combine stimuli but solely a sense (the simulated DA signal) that evaluates the choice and performance of every behaviour. In order to test this approach, the experiments in this chapter have addressed: the integration of external and internal stimuli in a variety of situations, the quickness of the learning process, and the quality of the adaptation via the behavioural analysis of the patterns learnt by the agent. This analysis initiates with an architecture consisting of three homeostatic variables, three drives and three behaviours; each of them related in a one to one fashion, e.g., the homeostatic variable nutrition expresses its status of deficit or excess via the drive hunger; both are compensated via executing the behaviour to eat. The actor-critic has demonstrated for both a stimulus driven and a motivation driven environment, that appropriate behavioural patterns can be learnt by interacting with the environment in a trial and error manner to maintain homeostasis (keep the physiological variables within their optimal values). This same analysis has been extended to two particular cases. Firstly, to an agent endowed with a new behaviour: to drink, the execution of which compensates several homeostatic variables. Secondly to the integration of appetitive and consummatory behaviours in a common competition for the agent's resources. For the aforementioned situations, the actor-critic and the learning paradigm have demonstrated that they are able to provide behavioural patterns that adapt the agent to this scenario, and to do so in such a manner that the agent's internal physiology remains stable.

The metrics and the presentation of results have been chosen to capture significant elements in the adaptation process to make them comparable to the methods used in ethology. In this respect, it is clear that the biological inspiration used to design the model does not suffice to explain several ethological phenomena; e.g., the combination of several appetitive behaviours in a single competition process is still limited. In this respect, it is important to highlight that this is a basic architecture aimed at understanding the way stimuli combine for a robotic application. Therefore, a single layered architectural behaviour has been considered for this aim. Nevertheless, the conclusions reached within this chapter suggest that this same architecture could be extended to build a multi-layered behavioural architecture where behaviours can self-organise according to their similarities, assessed by the environment, by using the same

learning paradigm.

The model introduced is inspired after the principle of ecology and after the learning hypothesis of the DA neurotransmitter as a effective reinforcement signal. These, supported by the results, suggest that learning and decision making are two complementary processes, ingrained in the hierarchy of adaption. Consequently, this leads one to view the several sorts of learning commonly classified as different processes in the psychological literature as particular cases of learning in a framework of ecological adaptation. The experiments have demonstrated that the learning process is a balancing process relating both sorts of stimuli with a single goal: homeostasis, hence maintaining the internal physiology within the viability range. Our main contribution to this understanding consists of viewing this process as a single dynamical process, where the dynamics of perception, internal physiology and interaction are sub-dynamics interacting with one another. Within this framework, this chapter has studied the effect of a variety of environments, demonstrating that the actor-critic can adapt to abundant and scarce environments as long as interaction is frequent enough to guarantee that its learning procedures can modify the behavioural patterns to survive. At a neurological level, these results suggest that the part of the brain participating in Pavlovian learning may also actively participate in the arbitration of motivation-mediated behaviours. However, this conclusion does extend to reflex movements or non-voluntary actions, since I argue that their execution is not based on reward.

Finally, it is argued that relevance of stimuli (measured in terms of reward) and habituation are fundamental to providing a sensible explanation for the motives that integrate external and internal stimuli in the way nature does. The learning process in this chapter has been mostly studied from the perspective of the environment. What changes in the agent's behavioural patterns are provoked by a change of the environment? However, there is a whole other set of elements affecting this process, the dynamics of the same agent's internal physiology. These are elements addressed in the next experimental chapter.

Chapter 6

Internal Modulation of Behavioural Patterns

The previous chapter has addressed the influence of the distribution and availability of environmental affordances in the learning of the actor-critic algorithm. The behavioural patterns arising from the interaction of the perception of external stimuli with the internal bodily dynamics have been assessed by measuring the physiological stability reached by each of them. These experiments have demonstrated that for the given environments, the actor-critic is capable of learning behavioural patterns and therefore capable of adapting on demand to the environment in a manner that fits ethological data. However, some patterns of animal behaviour cannot be understood if the environment is the only parameter considered experimentally. The behaviours of an animal exhibit large patterns of variation which can be better explained if both the internal and the external stimuli are considered as sources of adaptation. Therefore, in a complementary fashion to the preceding, this chapter focuses on the dynamics of the agent's internal physiology as a bias for adaptation. This process has been studied in the framework of a model, built to provide experimental data that help our understanding of how adaptive behavioural patterns emerge out of the combination of external and internal stimuli.

In an analogous fashion to previous ethological models, which were tested on robotic platforms (Spier and McFarland, 1996), I have embedded ours in a simulated robot for testing purposes. However, unlike for its previous counterparts, the goal of the experiments presented here is the study of the influence of the dynamics of the agent's internal physiology on its behavioural patterns. This relationship is one of the fundamental processes controlling the relationship of any animal with its niche.

The motivational model introduced in the previous chapter (see section 6.1) integrates internal and external stimuli in a single motivational state (Toates, 1986), which is used for learning and decision making in the context of an actor-critic. Previous experiments have indirectly tackled a qualitative idea of the influence of the internal physiological dynamics on the learning

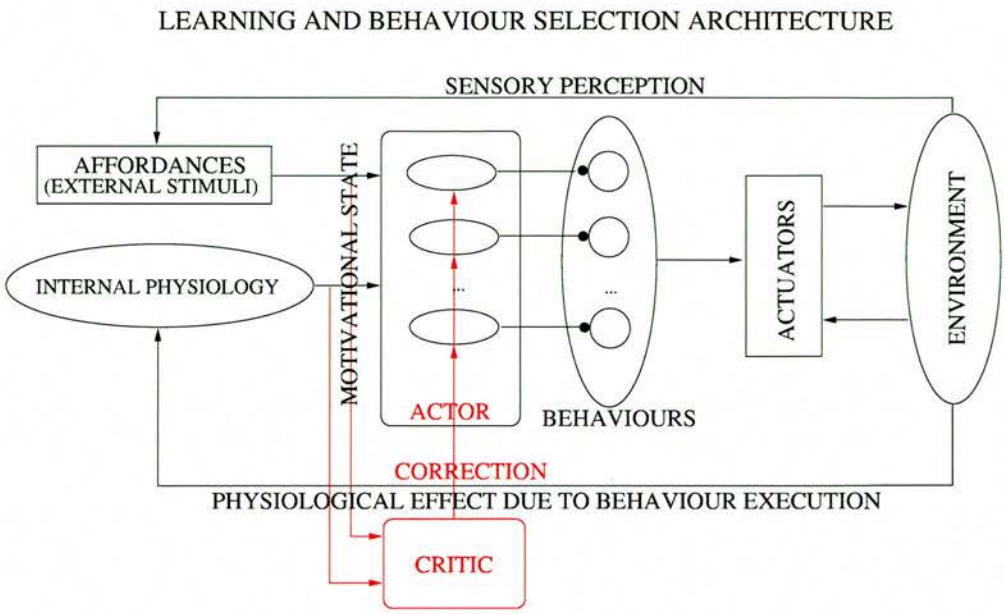


Figure 6.1: Architecture for Behaviour Selection and Learning. It consists of an internal physiology, a module to learn affordances, a behaviour repertoire and a set of actuators.

speed and on the behavioural patterns, but do not provide a quantitative account of these effects. This chapter aims at covering these deficiencies by studying a set of representative cases of modulation of behavioural patterns by the internal physiological dynamics. At this moment, it is important to note that among the different sorts of feedback provided by the environment, only reward has been considered. The notion of reward has been broadly addressed in the previous chapters. However, it is important to point out that reward is is in fact the simplest and most widespread sort of feedback in the animal world; animals ranging from a simple amoeba to higher vertebrates experience some type of feedback, which can be assimilated to some sort of reward, refer to section 4.2. Formally, I have adopted the following definition of reward: *"an operational concept for discovering the positive and negative values that a creature ascribes to an object, a behavioural act or an internal physiological state"* (Schultz et al., 1997). I argue that this value association is the cause of the modification of high-level goal-directed behavioural patterns, hence a method to improve the agent's adaptation. This argument is analysed in this chapter in the context of mobile robotics by considering the ecological principle as a boundary for acceptable solutions.

The chapter is organised as follows: the next section introduces the elements of the model concerning the characterisation of the agent's internal physiology and the context of study. This is followed by a description of the experimental setup preceding the experimental section.

6.1 Background Considerations

6.1.1 Modelling Internal Physiology

This section reviews the concepts of internal physiology with special emphasis on their influence on the learning and behaviour selection processes. The reader can refer to chapter 3 for a general overview of the model, to chapter 4 for further information about the manner in which perception has been modelled and to chapter 5 for a description of the learning and behaviour selection model, respectively. In a complementary fashion, this section will solely focus on the internal elements of modulation of the behavioural patterns, the physiological space (McFarland and Sibly, 1975), consisting of the homeostatic variables, their related drives and on their integration with the rest of the model.

Figure 6.1 shows an overview of the model. The homeostatic variables and their related drives are included in the *internal physiology* module. These variables represent the agent's internal resources, governed by the following dynamics (I repeat the equation from chapter 4 for convenience):

$$\tau_i \dot{V}_i = -V_i + \sum_j \alpha_{ik} \delta(t - t_j), \tag{6.1}$$

where V_i is the value of the homeostatic variable i and τ_i its related decay constant. The value of the variable is also affected by the execution of behaviours at times t_j. In case of a behaviour being successful, it provokes an instantaneous rise of α_{ik}, k is the index of behaviour b_k and i the index of the i^{th} homeostatic variable. The homeostatic variables express their status of deficit or excess through one or more related drives, as shown by

$$D_i = \sum_{k=1}^{N} a_k (V_{k_{op}} - V_k) + \sum_{j=1}^{N} b_j \dot{V}_j, \tag{6.2}$$

in its most general case. $V_{k_{op}}$ is the optimal value of the k^{th} related homeostatic variable. This relationship has been simplified in order to facilitate the analysis by associating every drive to a single homeostatic variable (there is a single index k in the first addition) and the dependence on the derivative of the homeostatic variable has been suppressed ($b_j = 0 \forall j$).

$$D_i = a_i (V_{i_{op}} - V_i), \tag{6.3}$$

which can also be expressed as

$$D_{i_k} = a_i ((V_{i_{k-1}} - (1/\tau_i) V_{i_{k-1}}) - V_{i_{op}}), \tag{6.4}$$

in its discretised form for simulation. Therefore, in this model *the regulation of each of the agent's drives depends on two parameters*: the V_{op} (optimal value or variable's set point) and the τ (decay constant) of its related variable. The system is endowed with three homeostatic variables: nutrition, stamina and curiosity, related one-to-one to the following drives: hunger,

tiredness and restlessness, respectively. The next section introduces two cases for the study of the influence of the internal stimuli in the process of behaviour selection and learning: stimulus driven and motivation driven.

6.1.2 Stimulus vs. Motivation Driven Learning and Behaviour Selection

The interaction between external and internal stimuli leads to behavioural patterns ranging from the stimulus driven (also called reactive) to the motivation driven, which are solely controlled by the agent's internal drives. Under normal circumstances, it seems reasonable that the resulting behavioural patterns would respond to a composite from both the demands of the internal physiology and the offers from the environment. In other words I adhere to the view that a good behaviour selection strategy should exhibit the right amounts of *persistence* and *opportunism* (McFarland and Spier, 1997; Tyrrell, 1993; Maes, 1991). Persistence is the capacity of continuing the execution of a behaviour the right amount of time and opportunism is the capacity of profiting from external stimuli despite having to contradict internal physiological advice (the drives).

I *hypothesise* that *the actor-critic can naturally learn behavioural patterns exhibiting the right amounts of persistence and opportunism*. The actor-critic maximises reward, in terms of which it has to learn to maintain the balance between stimulus and motivation driven behaviour selection strategies. In other words, this means persisting in the execution of a behaviour until its related homeostatic variable reaches its set point and to respond to the object affordances offered by the environment. However, following the tendency suggested by internal and external stimuli will often lead to incompatibilities. In these terms, it seems reasonable that *opportunistic behaviour* should dominate when the resources in the environment are scarce compared to the decay rate of its internal variables (see section 5.4). For example, a high need for food will be continuously expressed in the motivational state, thus biasing the selection of behaviours contributing to palliate this need when the environment affords to do so to the agent. These are demanding situations to address if the environment is not abundant in every sort of resource required by the agent. This can be perceived if I consider that the agent has often a single choice, the behaviour afforded by the stimulus. Therefore, if the encountered stimuli do not afford the behaviour to compensate the drive exhibiting the highest value, this may lead to values of the viability indicators, which could rise over the agent's lethal boundaries. Not for nothing did these situations exhibit the largest Risk of Death (RoD) (Ávila-García and Cañamero, 2004).

The actor-critic is hypothesised to provide patterns exhibiting opportunism and persistence that maximise reward by balancing the effect of the environment and of the internal physiological dynamics. If experiments in the previous chapter focused on the effect provoked by changes in the environment, this concentrates on the effect due to fluctuations of the agent's internal physiology at a behavioural level. This has been partly addressed by studying the

effect of the environment on the behavioural patterns in the previous chapter. In a complementary fashion, the experiments shown in this chapter address a quantification of the relationship between the agent's physiological dynamics and the behavioural patterns in two different scenarios. The first scenario represents an abundant environment where every object affords each behaviour to the agent. The second is a scarce environment only containing objects affording a single behaviour. As explained above, these two environments should give rise to reactive and motivation driven behavioural patterns, which are two extreme cases in terms of their distributions of resources, and should also lead to radically different strategies for behaviour selection. Furthermore, as a conclusion to the chapter, it also shows some experiments to survey the influence of the amount of compensatory effect on the homeostatic variables and therefore on the agent's adaptation. The next section introduces the experiments studying the effect of the dynamics of the agent's internal physiology on the learning and behaviour selection processes.

6.2 Experiments

6.2.1 Experimental Setup

The goal of the experiments is to analyse the influence of the internal bodily dynamics in the process of learning and behaviour selection. To this end, a simulated environment containing 10 objects has been engineered. Object encounters have been also simulated to minimise the simulation time. Therefore, objects are presented at random to the agent with equal frequency, and this executes a behaviour in a simulated time every time this happens. The agent's goal is to learn to keep its internal physiology stable, i.e., to reach its optimal physiological zone. To do so, the agent learns to execute sequences of behaviours in cycles, starting at a random point of its physiological space and finishing when its optimal physiological zone has been reached. At this moment the next cycle is started by resetting each homeostatic variable. It is expected that the learning process should statistically reduce the length of the cycle until reaching an optimal value.

The results have been analysed from a dual perspective. On the one hand, the internal physiology is monitored via the viability indicators; physiological stability and overall comfort (see equations 4.14 and 4.15). On the other, I have drawn the cycles of execution of behaviours and the measure of *effectiveness* during the execution of behaviours (see equation 5.8) for the different decay values (τ) considered. In order to isolate the effect of the dynamics of the internal physiology on the agent's behavioural responses, the amount of compensation α_{ik} caused by the execution of behaviour b_k (if it is successful) on variable V_i has been fixed to 0.3, being the only parametrisation in addition to the variable's decay constant τ_i. This latter parameter rules the pace of consumption of the agent's internal resources and should therefore influence the agent to interact to the environment. These effects are studied in two different

niches; in an *abundant* and in a *scarce environment*. These are introduced in the next two subsections.

6.2.2 Motivation Driven Opportunistic Agents

This section studies the influence of the agent's physiology on the learning and selection patterns for the case of an *abundant environment*. In this niche, every object affords every behaviour to be performed to our agent. This distribution of affordances is fixed throughout every simulation presented in this section; however, in order to *test the influence of the internal physiological parameters on the learning* process, two different physiological situations have been considered. Firstly, three sets of trials, where each homeostatic variable assumes the same τ value: 3×10^{-3}, 10^{-3} and 10^{-4} for every set of trials, respectively. Secondly, the fourth set of experiments will test the effect of having different decay constants for each homeostatic variable.

For the cases of having *homogeneous decay constant values*, figures 6.2 and 6.3 show the evolution of the mean cycle length and of the evolution of the viability parameters, respectively. The evolution of the cycle length over the simulation (cf. figure 6.2) shows that the learning response depends on the decay of the homeostatic variables (τ) for the values of $\tau = 3 \times 10^{-3}$, $\tau = 10^{-3}$ and $\tau = 10^{-4}$.

The larger its value, the longer needs the robot to minimise the length of the cycle. The final mean number of decisions varies between 5 and 7 and its variance between 2 and 3, depending on the decay constant. Furthermore, I can guarantee that for the given scenario any decision has an effect on the internal physiological dynamics (unless the satiation boundary has been reached) since any object affords every behaviour to be executed. Furthermore, these viability values are the lowest boundary values that can be reached by this agent provided that the amount of compensation by any behaviour is fixed at α_{ik}=0.3. Furthermore, it has been experimentally demonstrated that the actor-critic diverges for τ values larger than 3×10^{-3} (for the given α_{ik} value). When all three homeostatic variables decay this fast, the compensatory effect due to behaviour's execution is not sufficient to compensate their related deficits, which leads to the death of the agent.

In a similar fashion, figure 6.3 shows that τ has a similar effect on both the physiological stability and the overall comfort. These indicators improve (diminish their values) when the decay constant exhibits its lowest values. Their stationary values exhibit the same pattern. The stability ranges lower than 0.1, and the overall comfort under 0.005. These are the lowest values that can be reached for the given experimental setup.

However, in order to reach conclusions about the influence of the physiology it is necessary to analyse the resulting behavioural patterns. To this end, figure 6.4 compares the matching between the learnt patterns and purely motivation driven ones. The graphs obtained demonstrate

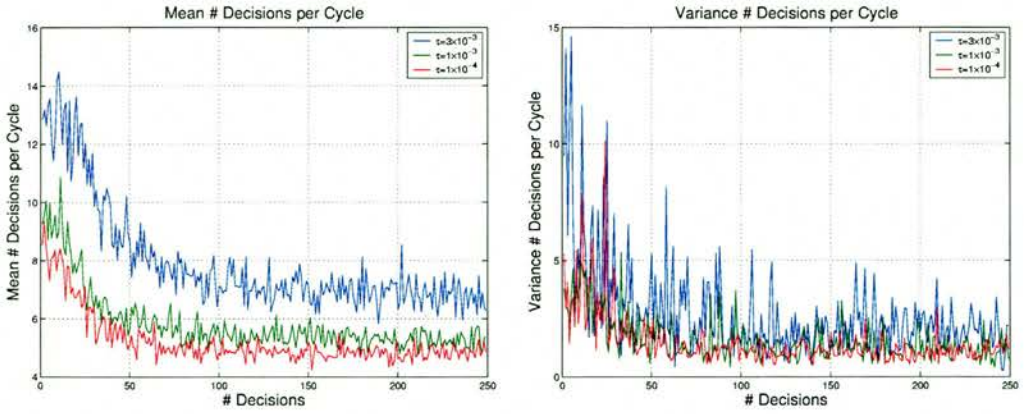


Figure 6.2: Mean length of the behavioural cycle for the case of the abundant behaviour, mean and variance, left and right, respectively. The x-axis is time, measured in number of decisions. The tick values are divided by 2000 (multiply by 2000 to get the actual number of decisions). The y-axes is the length of the learning cycle.

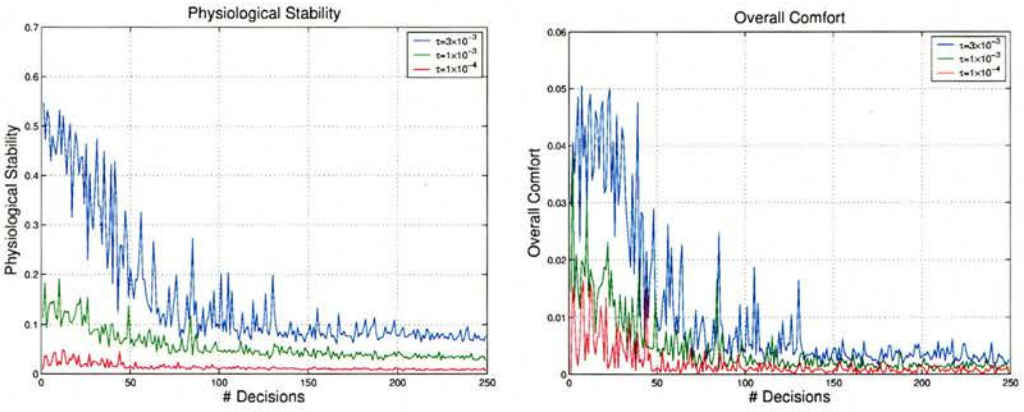


Figure 6.3: Effect of Internal Modulation on the Evolution of the Viability Indicators, physiological stability and overall comfort, left and right, respectively. The x-axis is time, measured in number of decisions. The tick values are normalised by a factor of 2000. The tick values are divided by 2000 (multiply by 2000 to get the actual number of decisions). The y-axis are the physiological stability and overall comfort, left and right, respectively.

that for the case of this abundant environment and for the given decay constants and α_{ik} parameters, the learnt patterns are close to being exclusively driven by the agent's physiology. To understand how the calculations have been performed, it is necessary to consider that in this architecture and experimental setup every behaviour is related one-to-one to a drive. In other words, this graph shows the percentage of matching between the behaviour selected at time t and the behaviour related to the drive exhibiting its highest value at time t (the one expressing the highest urgency). At the end of the simulation the percentage of matching is around 90%

which means that the agent's internal drives are controlling the strategy to select behaviours. Informally, the agent has learnt to serve its most urgent drive. This pattern is consistent if I consider that any object in this abundant environment affords every behaviour to the agent. Therefore, any decision that matches the agent's internal needs (motivation driven) will be rewarding independently of the sensory input.

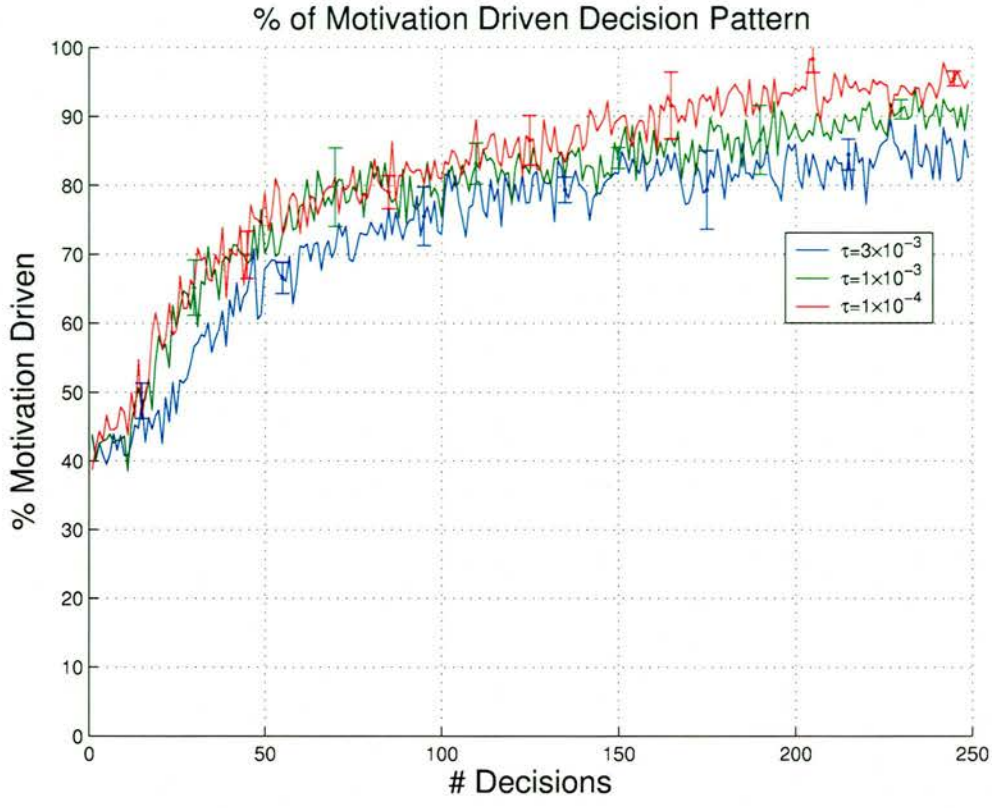


Figure 6.4: Evolution of the percentage of motivation-driven decision pattern. Since each behaviour is related one-to-one to a drive, this indicates the percentage of decisions where the behaviour selected corresponds to the drive exhibiting the highest urgency, therefore the percentage of decisions driven by the internal drive values.

The first three sets of experiments have demonstrated that the internal physiology in this environment has a clear effect on the learning performance and that the learnt behavioural patterns are primarily motivation driven if the resources are abundant in their scenario. However, more experiments have to be performed in order to assess the types of responses that the internal physiology provokes at a behavioural level. To this aim, a fourth set of experiments has been used for comparison with the former. In this case the decay constant of each homeostatic variable is different: the level of nutrition of the first agent decreases very fast ($\tau = 3 \times 10^{-3}$), its level of stamina decreases at an intermediate pace ($\tau = 10^{-3}$) and its level of restlessness decreases very slowly ($\tau = 10^{-4}$). Does the agent serve more often the behaviour whose related

homeostatic variable decays faster? The evolution of the mean length of the cycle of execution to reach the viability zone and the evolution of the viability parameters are presented in figures 6.5 and 6.6, respectively.

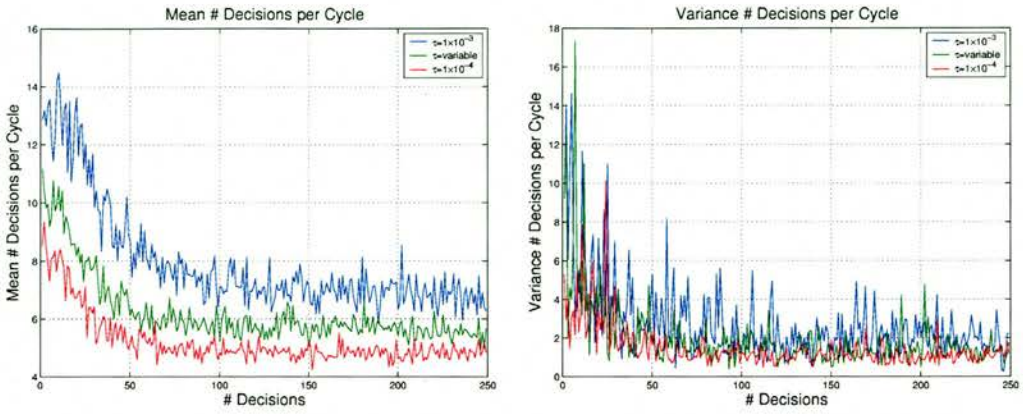


Figure 6.5: Mean length of the behavioural cycle for the case of the abundant environment, mean and variance, left and right, respectively. The x-axis is time, measured in number of decisions. The tick values are normalised by a factor of 2000. The tick values are divided by 2000 (multiply by 2000 to get the actual number of decisions). The y-axis is the length of the learning cycle.

Figure 6.5 shows the mean and the variance of the cycle length averaged over 20 simulations. The results obtained from this fourth set of experiments are compared to two of the former cases (symmetric $\tau = 10^{-3}$ and $\tau = 10^{-4}$). It shows that the results for this case lie between the two former cases.

Figure 6.6 shows the evolution of the viability indicators, physiological balance and overall comfort, for this fourth set of experiments. The values are compared to two of the simulation sets performed with an agent endowed with the same τ decay constant for every one of its homeostatic variables, $\tau = 3 \times 10^{-3}$ and $\tau = 10^{-4}$, respectively. Consistently with the graphs in figure 6.5, the viability indicators improve throughout the simulation in a manner that averages both the former cases. This is due to the fact that the mean decay value of the three asymmetric τ values lies between the two symmetric τ values, which suggests that the actor-critic is learning patterns which are an average of the patterns exhibited in the two cases used for comparison. To confirm this, it is necessary to observe the behavioural patterns obtained in both cases. These are shown in figure 6.7. On the left hand side, the case of having the three symmetric τ values equal to 10^{-3}, on the right hand side the aforementioned asymmetric distribution of τ.

Cycles of Execution The pattern has been executed by starting at the point (0.1, 0.5, 0.5), where each coordinate stands for the hunger, tiredness and restlessness drive values, respec-

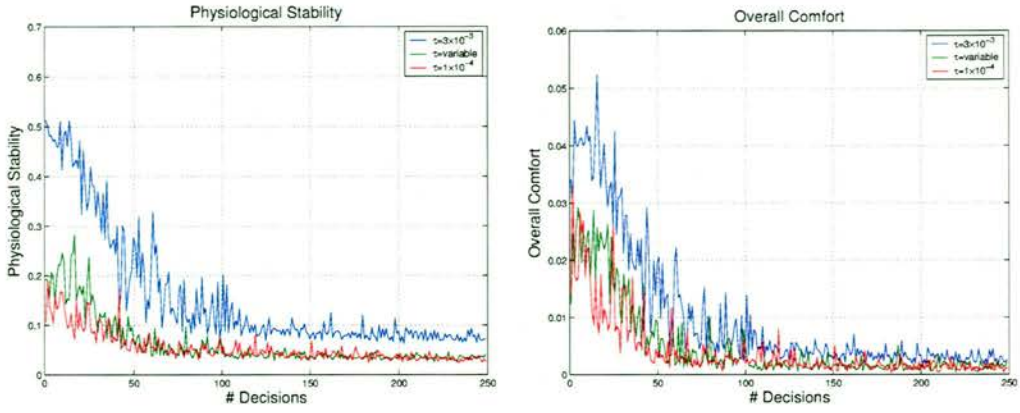


Figure 6.6: Effect of Internal Modulation on the Evolution of the Viability Indicators, physiological stability and overall comfort, left and right, respectively. The x-axis is time, measured in number of decisions. The tick values are normalised by a factor of 2000. The tick values are divided by 2000 (multiply by 2000 to get the actual number of decisions). The y-axis are the physiological stability and overall comfort, left and right, respectively.

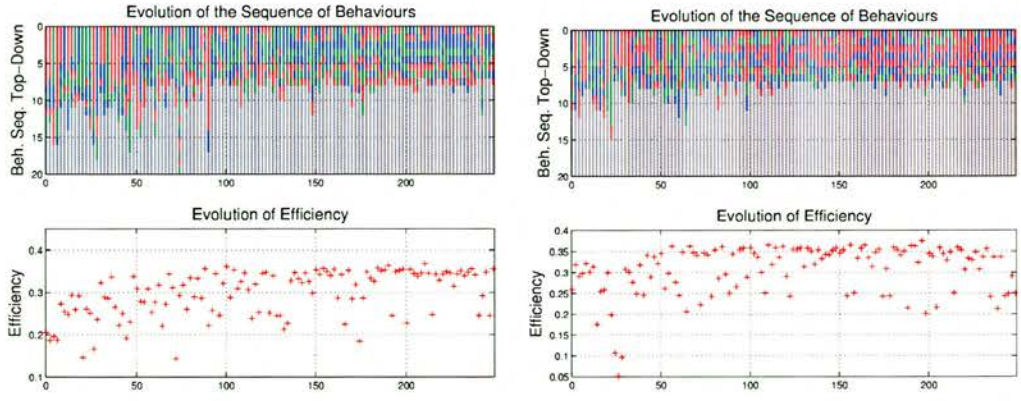


Figure 6.7: The top graphs represent the evolution of the behavioural cycles for the case of $\tau = 3 \times 10^{-3}$ for each variable and for the case of nutrition ($\tau = 3 \times 10^{-3}$), stamina ($\tau = 10^{-3}$) and restlessness ($\tau = 10^{-4}$)), left and right, respectively. The bottom graph shows the effectiveness of the behaviour execution. The behaviours are represented by colour: red (grasp), green (to shelter) and blue (touch).

tively. Therefore, every time the cycle starts, the agent experiences a high need for nutritional supplements as expressed by the value of its related drive. This reflects on the behavioural patterns independently of the agent's internal physiological parameters. The pattern shows the evolution of the sequence of execution of behaviours from beginning to the end of the simulation. Each cycle is represented by a vertical line flowing from top to bottom, where the execution of every behaviour is represented as a dot, in red, green and blue for the behaviours

eat, rest and touch, respectively. The patterns in figure 6.7 show that the pattern becomes shorter after 150 iterations. This means that the behavioural pattern is starting to be *effective* (as expressed by the red dots drawn below —their optimal value is 0.4 and for both cases, the final value reached is about 0.35). The influence of the internal physiology is noticeably. The cycle on the right hand side exhibits a lot more red colour (more executions of the grasping behaviour), which means that an agent becoming hungrier at a very fast pace also learns to compensate this need by executing the behaviour grasping more frequently. In terms of persistence, it can be argued that the execution of a behaviour is maintained until its physiological needs are covered.

These results, obtained in four different cases, demonstrate that the internal physiological dynamics do exert an influence on the learning process and on the final learnt behavioural patterns. However, this holds when the sensory input carries no information, since the niche has been designed in a manner that any object affords to do anything to the agent. Therefore, it could be expected that the internal physiological dynamics do influence the agent's learning and decision processes. In order to complete this view, the opposite case of niche has been considered, a second environment with a particular distribution of affordances. In this case the sensory input does carry some information that the agent has to learn to consider to make decisions. However, what is the influence of the internal dynamics in this case? This is addressed in the next section.

6.3 Stimulus Driven Behaviour

To complete the comparative analysis introduced in the experiments of the previous section, the experiments presented here are performed in an environment where each object affords a single behaviour to be executed. As demonstrated in the previous chapter, the learning of behavioural patterns for such environments turns out to be more difficult, since the only reasonable choice in terms of reward consists of reacting to the offered affordance (see section 5.4). The distribution of affordances used for the experiments in this section is shown in figure 6.8.

Therefore, if the behavioural pattern is correctly learnt, the expected response should ignore the internal physiological dynamics independently of its τ decay values. To test this, three sets of experiments have been parametrised after τ. In each set of experiments the three decay constants of the three homeostatic variables have been endowed with the same value. The τ values for each set of experiments have been 2×10^{-3}, 10^{-3} and 10^{-4}.

Figure 6.9 shows the evolution of the length of the cycle obtained by averaging 10 simulations for each τ value. These results confirm that the cycle grows longer when the τ becomes larger. Furthermore, it also takes longer to learn for high τ values. It is important to notice that patterns to maximise reward arise in this environment only when the decay constant is smaller

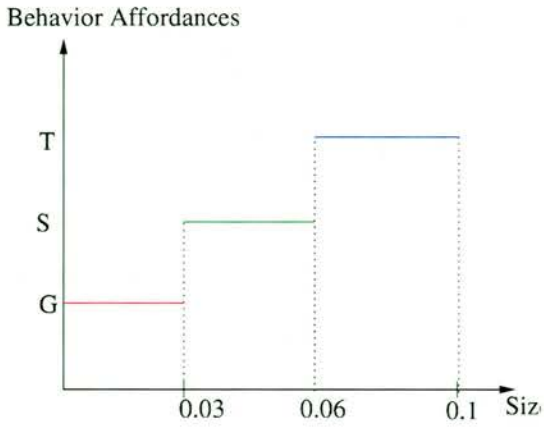


Figure 6.8: Scarce Distribution of affordances. The line indicates the interval of object size where that affordance is 1.0. G, S and T stand for grasping, shelter and interact, respectively. Object sizes range from 0.0 to 0.1.

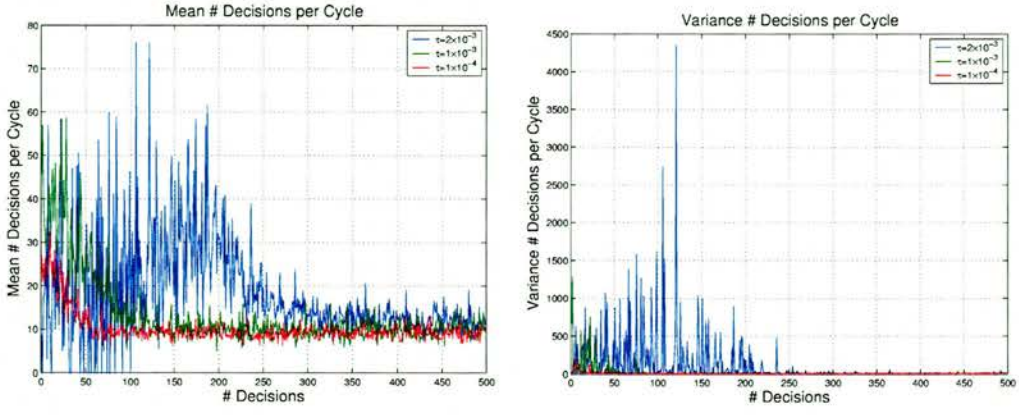


Figure 6.9: Mean length of the behavioural cycle for the case of the abundant environment, mean and variance, left and right, respectively. The x-axis is time, measured in number of decisions. The tick values are normalised by a factor of 2000. The tick values are divided by 2000 (multiply by 2000 to get the actual number of decisions). The y-axis is the length of the learning cycle.

than $\tau = 2 \times 10^{-3}$. The evolution of the cycle length for this τ value is plotted in blue, which already exhibits a noticeable longer time until stationarity and strong oscillations during the first half of the simulation. The main reason for the reduction of the range of τ values arises from the fact that this environment is far more hostile than the abundant environment. In this case it does not permit the agent to perform the behaviour it needs to compensate its most urgent drive. Instead, it is forcing the agent to, at most, react to the offered affordances in a manner that may not be appropriate to the agent's needs.

Similar conclusions can be drawn from the viability indicators shown in figure 6.10. Again,

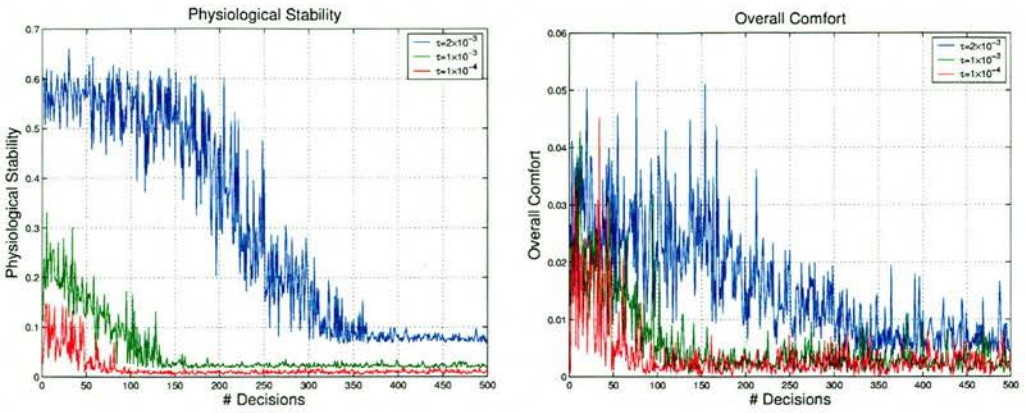


Figure 6.10: Effect of Internal Modulation on the Evolution of the Viability Indicators, physiological stability and overall comfort, left and right, respectively. These values have been measured for the case of a scarce environment. The x-axis is time, measured in number of decisions. The tick values are normalised by a factor of 2000. The tick values are divided by 2000 (multiply by 2000 to get the actual number of decisions). The y-axis are the physiological stability and overall comfort, left and right, respectively.

these graphs are the result of averaging over 10 simulations each. The graph of the physiological stability shows that optimal (minimal) values are reached when the cycle of execution reaches its shortest length. Their stationary values are for all three cases under 0.1 (10% of the range of the drives). The graph on the right hand side indicates the overall comfort, which again exhibits how the deficits vary throughout simulation. Despite having a rough start, the final stationary values lie within 0.01, which can be considered as extremely good (less than 1% of the range of the drives). These figures together with figure 6.9 suggest that the dynamics of learning process and of the viability indicators are intrinsically related. However, reaching further conclusions requires some further exploration.

To this end, I have *analysed the patterns of behaviour selection* by comparing them with two extreme behavioural cases: a purely stimulus driven behavioural pattern and a purely motivation driven behavioural pattern. The method applied consists of comparing every decision of the pattern learnt by the agent with a reactive and a motivation driven pattern. Figure 6.11 shows these comparisons, left and right, respectively.

These show that the internal physiological dynamics has severe implications on the learning process, as depending on the internal rhythm, the behavioural pattern resembles more a reactive pattern or a motivation-driven pattern. For *large* τ values, such as 2×10^{-3} and 10^{-3}, the influence of the stimulus grows over time and is maintained throughout the entire simulation. However, for the smallest value 10^{-4} (line in red) it consistently grows until reaching 80% of the stimulus-driven pattern and then decreases until 50%. I suggest that there is a double

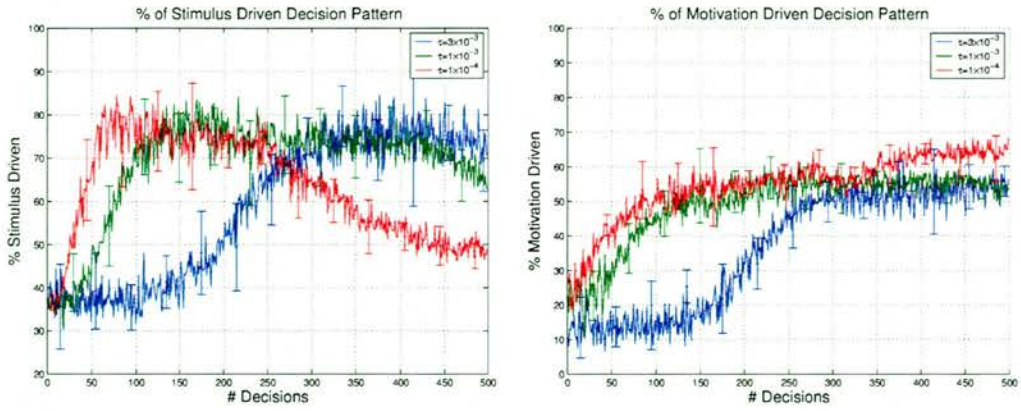


Figure 6.11: Percentage of matching of matching between the learnt behavioural patterns in a scarce environment and the purely stimulus or motivation driven behavioural patterns, left and right, respectively. These values are compared for three different values of τ $\tau = 3 \times 10^{-3}$, $\tau = 1 \times 10^{-3}$, $\tau = 10^{-4}$.

effect. Firstly the patterns are learnt, however, the more efficient these are, the more sated is the agent and therefore, the least needed is that efficiency in selecting the right behaviour. Since τ is very small, the reward that results from executing the reactive behaviour suggested by the stimulus may not be incentive enough for its execution. For the case of very slight physiological decays, it may not be necessary to execute a consummatory behaviour at every opportunity, hence choosing instead any behaviour may not have a negative effect in this particular case. The difference between left and right graphs may suggest that a loss in reactivity turns into a gain of drive-driven behavioural pattern. However, if our previous assumption is correct, this conclusion would only be partly right, since both the gain in drive influence and the loss of affordance influence would be a result of the lack of necessity of the agent to perform a behaviour, therefore just a consequence of physiological constant decays that are very slow. In a way, this also suggests that the agent has learnt to balance the opportunism and persistence. Opportunism to react appropriately to the given stimuli and persistence to continue executing the same pattern, since it leads to more stable physiological values (as shown in figure 6.10).

These conclusions, however, have been reached only for the case of a symmetric distribution of decay values. To extend this to a more general case, a fourth set of experiments has been performed where each homeostatic variable is endowed with a different τ decay value. For this case, the level of nutrition of the first agent decreases very fast (2×10^{-3}), its level of stamina decreases at an intermediate pace (10^{-3}) and its level of restlessness decreases the slowest (10^{-4}).

The final behavioural patterns exhibit again an appropriate level of persistence and of opportunism. Figure 6.12 shows the evolution of the mean length of the cycle of execution from

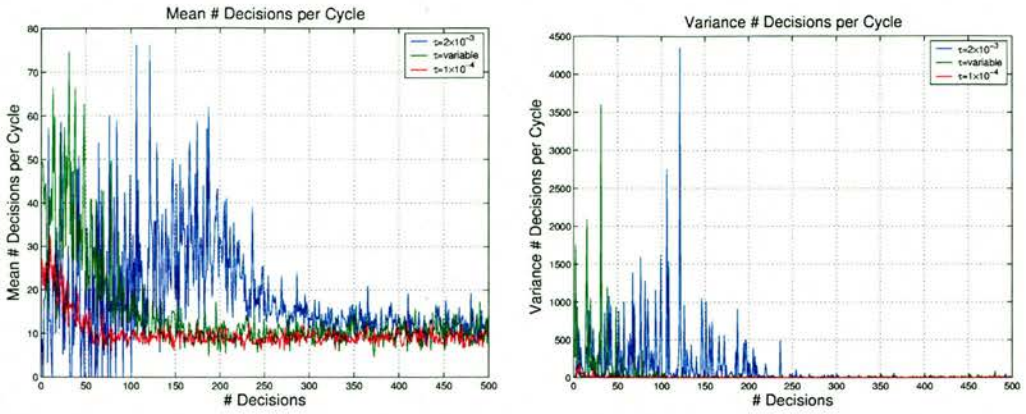


Figure 6.12: Mean length of the behavioural cycle for the case of the abundant environment, mean and variance, left and right, respectively. The x-axis is time, measured in number of decisions. The tick values are normalised by a factor of 2000. The tick values are divided by 2000 (multiply by 2000 to get the actual number of decisions). The y-axis is the length of the learning cycle.

a random starting point to the agent's optimal zone (each curve has been obtained by averaging over 10 simulations) for the case of asymmetric decay constants (in green), and compares its behaviour with the cases of $\tau = 2 \times 10^{-3}$ and $\tau = 10^{-4}$. The green line shows that the length of the cycle decreases until it reaches a stationary value, which is nearly independent of the decay constant. This is important, since when I pay attention to the viability indicators, cf. figure 6.13, there are significant differences between the three cases, laying the case of asymmetric decay constants between the two boundary cases (the graphs are obtained by averaging over 10 simulations). In fact, the smaller the decay constant of the related homeostatic variable the smaller the values of physiological stability and overall comfort.

In order to ground this difference, it is necessary to pay some attention to the *behavioural patterns* studied in the three preceding cases, where the τ was the same for each homeostatic variable. When comparing the behavioural patterns in figure 6.11, I observe a shift of the behavioural patterns from more stimulus driven to more motivation driven. This change is probably due to *satiation*; the better the patterns, the more reward it will get and the more stable will be the agent's physiology. Therefore, it is likely that every homeostatic variable will from then onwards be constantly sated. In such a state there is no difference in terms of physiological stability and it can be argued that the performance of any behaviour, either more stimulus driven or more motivation driven, has no effect either in terms of stability or in terms of reward, since there is no need for a particular behaviour's execution.

Figure 6.14 illustrates the behavioural patterns obtained in simulation for two different cases. For the first one, its homeostatic variables decay with the same τ value equal to 10^{-3};

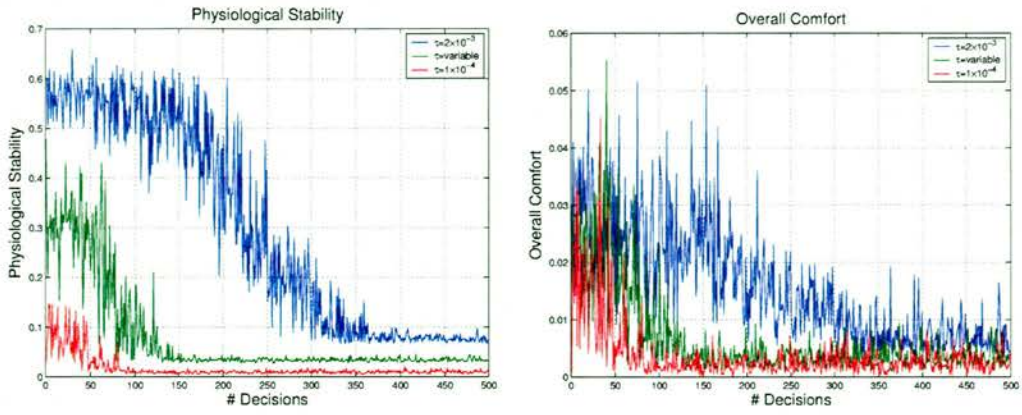


Figure 6.13: Effect of Internal Modulation on the Evolution of the Viability Indicators, physiological stability and overall comfort, left and right, respectively. The x-axis are the number of decisions considered analogously to time. The tick values are normalised by a factor of 2000. The y-axis are the physiological stability and overall comfort, left and right, respectively.

for the second agent, the variables are endowed with the asymmetric values introduced above.

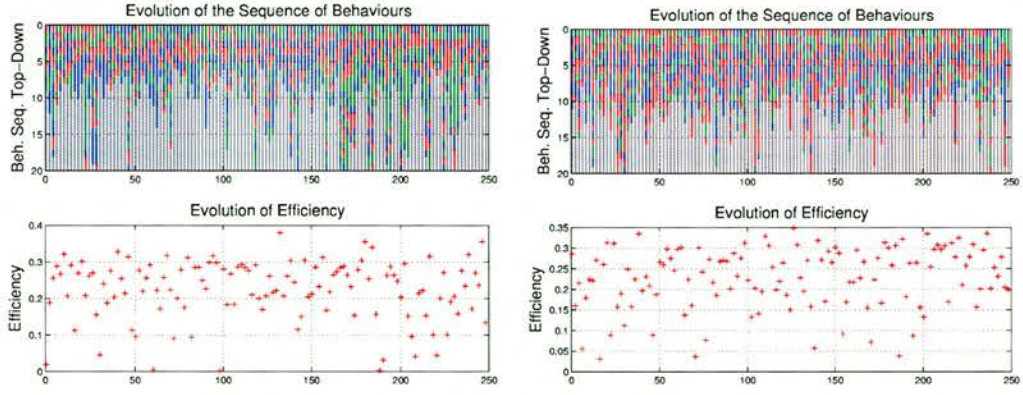


Figure 6.14: Evolution of the percentage of motivation-driven and stimulus-driven decision pattern. Since each behaviour is related one-to-one to a drive, and every affordance is by definition related to a single behaviour, these graphs show the percentages of decisions where the behaviour selected corresponds to the drive exhibiting the highest urgency (left) and the percentage of reactive behaviour (on the right hand side).

The patterns show cycles starting always at the same point of the physiological space: (0.1, 0.5, 0.5), the coordinates of which stand for hunger, tiredness and restlessness, respectively. Therefore, at the beginning of every cycle, the homeostatic variables are reset to the following values: nutrition to 0.9, stamina to 0.5 and boredom to 0.5. It can be observed that in the case of having different τ decay constants (cf. pattern on the left in figure 6.14), the agent learns to responds faster to those homeostatic variables experiencing a faster depletion. This is shown

by more executions of the behaviour shelter (green). The behaviour to grasp is executed less frequently than the rest, despite having a larger τ constant, this homeostatic variable is sated at the beginning of the cycle. It can be observed that the cycles at the end of the simulation rarely start with the execution of this behaviour; the few cycles in which this happens are due to random selection. Conversely, for the case of having homogeneous τ decay constants (cf. pattern on the right in figure 6.14), the patterns appear to contain frequent executions of the behaviours shelter and touch (green and blue), and slightly less executions of the behaviour grasp (red). This may suggest that effectively, the cycle tends to respond to the needs expressing the highest urge; in this case to shelter and interact.

These results demonstrate that the internal physiological dynamics do exert an influence during the learning process and on the final behavioural patterns. Furthermore, the effects observed in these results confirm that the dynamics of the agent and of the environment are part of a single dynamics, where needs and availability of resources influence the manner in which we interact with the environment. These also show that it is the combined influence of the external and internal stimuli that determines the patterns. It does not suffice to explain that the most restrictive of both stimuli, the external or the internal, drives the decision making (this could be explained via the $affordance \times drives$ multiplicative formula to calculate the intensity of the related motivations). However, this multiplicative rule does not explain stimulus driven behaviour (according to the formula, the intensity of this motivation should be zero if the drive is zero); a behaviour that the actor-critic has been demonstrated to be able to learn. This has been demonstrated to hold for the case of having a symmetric effect value (α_{ik}) for every behaviour, furthermore when the effect is related to the reward by the formula introduced in figure 5.2. This also demonstrates that both the environment and the dynamics of the agent's internal physiology influence the learning process and the resulting behavioural patterns. However, the actor-critic has shown that the resulting learnt behavioural patterns do exhibit an appropriate level of persistence and opportunism for the set of environments and for the physiology used for simulation.

Nevertheless, these results are a consequence of the assumption that behaviours contributing to the internal physiological stability are to be considered good and are therefore associated to a positive reward value. Otherwise the value is negative. This may not necessarily always be the case, since there is not a unique relationship between the effect of a behaviour and the reward I experience. This is introduced in the next section.

6.4 From Effect to Reward

The term *reward* has been so far used in an abstract sense that refers to the event consistently following the presentation of a stimulus. However, reward is a broad concept with significant

cognitive implications. Reward can also be referred to as the *valency* Ackley and Littman (1991), which is the affective interpretation of the physiological effect provoked by the execution of a behaviour. This sense may have arisen because it facilitates survival in a competitive niche. It is straightforward that animals that associate a good feeling to behaviours that improve their physiological state have a larger probability of surviving than others. This is an enhanced solution to the need of maintaining physiological stability proposed by Ashby (1965), which we have used as an ethological constraint. Therefore, *reward* can be modelled as an assessment of the physiological effect provoked by a behaviour. Through this perspective I have designed a reward formula inspired from these principles. It delivers reward for any decision leading towards physiological stability and a punishment otherwise. Among the infinite formulae that quantify behaviour and the aforementioned constraint I have chosen the following

$$r(t) = \beta[\frac{1}{\|\bar{d}_i(t)\|^2} - \frac{1}{\|\bar{d}_i(t-1)\|^2}], \tag{6.5}$$

where $\bar{d}_i(t)$ and $\bar{d}_i(t-1)$ are the current and the previous physiological states, respectively. This formula relates effects diminishing the deficits to a positive value, coherently with the aforementioned constraint (cf. figure 6.15). β is a scaling factor.

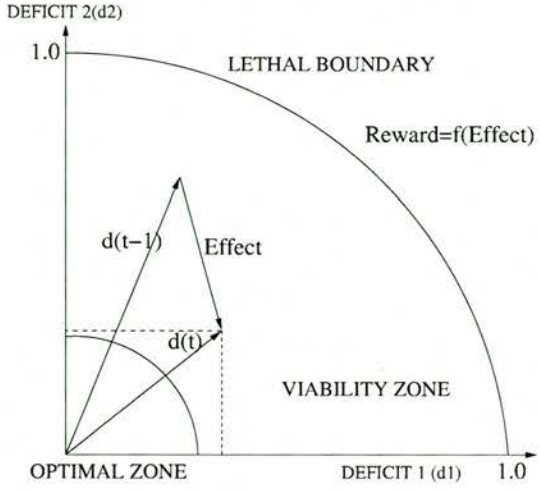


Figure 6.15: Definition of Reward in a 2D Physiological Space. $d(t-1)$ and $d(t)$ stand for the state before and after the execution of a behaviour. The vector of effect stands for the amount of effect due to the execution of a behaviour. The Optimal Zone represents the area of minimal deficits.

The hypothesis introduced by formula 6.5 is vital for several reasons. On the one hand it introduces the sufficient constraints to extend the learning hypothesis of Schultz et al. (1993) to instrumental learning (Houk et al., 1995), since now the delivery of reward is always mediated by the execution of the appropriate behaviour. On the other, this formula respects basic ethological constraints while not imposing any additive or multiplicative formulae to combine external

and internal stimuli to compute the motivational state. The only inherent condition imposed on the behavioural patterns of the algorithm is that these must maximise the reward within the cycle of execution. Reward-driven decision making has also been suggested in neuroscience (Rolls, 2003).

This relationship between reward and effect has been tested by running a series of simulations where several β values in equation 6.5 have been used to distort this relationship. The experimental setup used is introduced next.

6.4.1 Experimental Setup

For both experiment sets, the robot is placed in two sets of *environments* as described in the previous sections.

The same metrics on **learning velocity** and **physiological stability** introduced in section 5.3.3 have been again applied to the experiments of this section. The robot navigates at random. Every time an object is encountered, the state is updated by perceiving the set of external (*affordances*) of the object encountered and by reading the instantaneous value of the agent's internal drives (*drives*). The actor calculates then the motivational state (the policy values) and the behaviour whose related motivation exhibits the highest value is selected and executed. Then the object is abandoned, to wander at random until another object is encountered to restart the cycle of execution.

The goal of the experiment set is to *evaluate the influence of effect in the physiological stability*. The amount of effect and its consequent interpretation as reward during the learning process determine not only the pace of learning, but also the quality of the final values for convergence. To this aim, the effect of a behaviour (β parameter) has been parametrised for each behaviour between 0.15 and 0.35.

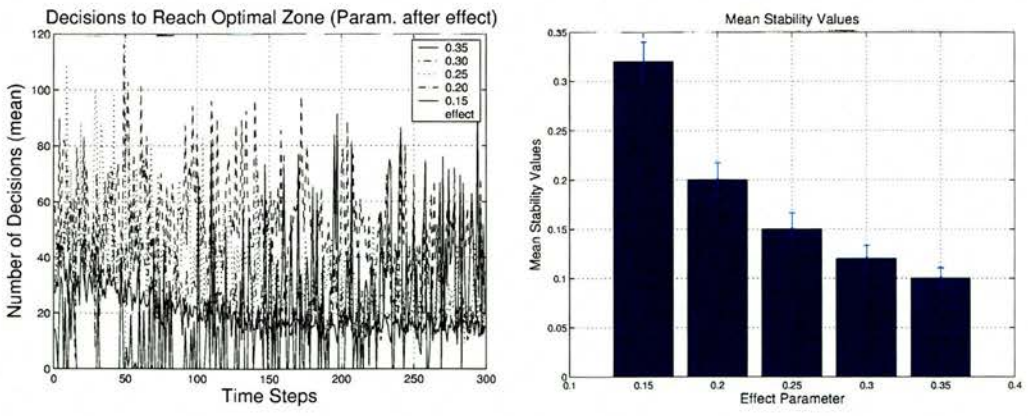


Figure 6.16: Decisions per Cycle and Physiological Stability for different amounts of effect, left and right, respectively.

The *results* shown in figure 6.16 show that the level of distortion is directly proportional to the values of stability obtained for simulation. The larger is the effect of each behaviour (its step value β), the shorter is the cycle of execution and the smaller is the stationary value of the physiological stability.

This has to be considered in the ecological framework that characterises the relationship between the agent, in terms of internal physiology, and the environment in terms of how this affects the agent. The more available are the resources in the environment, the easier it should be to learn policies to satisfy internal needs. Likewise, the larger is the effect with regard to the pace of growth of the deficits, the shorter is the cycle and the lower the mean of the deficits, cf. figure 6.16.

Conclusion

This chapter has addressed the study of the influence of the decay constants of the agent's homeostatic variables and of their effect on the learning process and on the behavioural patterns. The results have demonstrated that these two internal factors (the decay constants and the amount of effect due to every behaviour) condition the resulting behavioural patterns.

The *decay constants* of the homeostatic variables are demonstrated to influence the learning and behavioural patterns in a variety of manners. In order to study these effects, I have designed several sets of experiments considering the phenomena which depend on the environment (see chapter 5 for further detail). Experiments in chapter 5 have demonstrated that the responses of the actor-critic vary depending on the availability of resources of a scenario. Therefore, a scarce and an abundant environment have been engineered and used for simulation.

The experiments in the *abundant environment* have highlighted that the smaller the decreasing factor of the homeostatic variables, the longer the time needed to learn the optimal behavioural patterns and the worse the final viability values will be. However, these experiments have assumed that every single homeostatic variable exhibits the same decay rate. In terms of the actor-critic, this means that every homeostatic variable equally contributes to the reward signal used for training (cf. figure 5.2). However, this does not extend the palette of behavioural responses that the actor-critic can obtain under the effect of the internal physiology only. To cover this, a second set of experiments where every homeostatic variable is endowed with a different decay constant has been performed. The behavioural patterns obtained in this case demonstrate that the actor-critic has learnt to respond more frequently to the variables decreasing faster. These first experiments have highlighted that the τ decay constant has a strong influence on the learning and on the behavioural patterns. However, this covers so far only an environment where every single object affords every behaviour to be performed by the agent. Under these circumstances it seems natural that the decay constant will dominate the

actor-critic responses.

For this reason, a *scarce environment* has also been considered as a testbed. Objects in this environment afford a single action to be performed by the agent. This niche had been previously considered in chapter 5, where it demonstrated that the most efficient behavioural pattern in terms of reward was reactive behaviour; therefore a niche where the internal physiology seemed to have little or no influence. However, after running a series of experiments parametrised after the decay constant, it has been demonstrated that physiology does influence both the learning and the behavioural patterns also in this case. In particular, when the decay constant reaches very small values in proportion to the compensatory effect of every behaviour, the final behavioural patterns differ from a reactive pattern. In contrast, the final pattern becomes partly motivation driven and partly stimulus driven. I observe that the same viability values have been reached for different decay parameter values in the stationary regime. It is therefore straightforward to conclude that the homeostatic variables for small decay constants become easily sated and that several behavioural responses exhibit the same performance in terms of internal physiological stability. An explanation results if I consider that slow varying homeostatic variables tend to be sated when the behavioural executions are quicker than the internal needs can express. The satiation boundary is reached in this case, for which any variation in the behavioural patterns does not provide any better physiological stability, as shown by the aforementioned experiments. This has also been confirmed by the behavioural patterns shown in figure 6.14. The two patterns shown compare the case of three homeostatic variables endowed with the same decay value to a final case where every decay constant is different. The patterns show that there is not a constant pattern as happened in the abundant environment, since although every cycle starts with the same physiological deficits, the behaviour is mostly reactive. This is because it depends on the affordances of the objects encountered at random. However, since the τ decay constant of the homeostatic variable nutrition is larger than the rest, the behaviour grasping is executed more often than the rest. These experiments, performed in two environments endowed with very different distributions of resources, have demonstrated that the pace of decrease of the homeostatic variables does condition the learning process and the final behavioural patterns.

The influence of the internal physiology does not only depend on the decay constant of the homeostatic variables. The probability of an agent of persisting in one behaviour or another or of being more or less opportunistic (reactive) also depends on *reward* and how reward is defined. At the beginning of the chapter I argued that this is one of the simplest manners for the environment to affect the agent's internal stability. In an introductory fashion I have introduced a last set of experiments to measure the influence of the amount of effect due to the execution of behaviour on the agent's homeostatic variables. These have shown that the internal physiological stability depends on the amount of compensation provided by every behaviour.

In general terms, the larger the compensation the better the viability values.

With these sets of experiments, the actor-critic has been demonstrated to provide appropriate policies to maintain physiological stability in a variety of scenarios, with different availability and accessibility of resources and parametrisation of its internal physiology. This also suggests that reward is related to stability and that our formulation of reward is sufficient to provide explanation of a variety of ethologically resemblant patterns with a simple model. Furthermore, this also supports the hypothesis that dopamine (DA) in the basal ganglia is not only the error of the prediction of reward for Pavlovian, but also for Instrumental learning. Hence, it does not only learn to relate stimulus to responses, it may also be biasing the selection of one behaviour over another. This suggests, from a physiological perspective, that the parts of the brain involved in learning Pavlovian contingencies (as the experiments of Schultz demonstrate), would also be involved in instrumental learning.

Chapter 7

Discussion

This dissertation has addressed adaptation from the perspective of a situated, embodied agent (Agre and Chapman, 1987) by adhering to a view on intelligence based on the principle of interaction with the environment. This perspective of intelligence is partly motivated as a response to the limitations demonstrated by symbolic classic AI (Brooks, 1986). Clearly, despite novel and classic AI sharing the term intelligence, its semantics have a different flavour in each case (Steels, 1994). While symbolic AI has been mostly concerned with intelligence as a logical manipulation of beliefs and with faithful representations of the world (Gärdenfors, 1992; Harnad, 1990), novel AI is most concerned with intelligence in terms of effective interaction with the environment, where even a model of the environment is often not necessary. A natural consequence of this approach has been that the problems on which AI has primarily focused are relatively simple (nonetheless interesting) ones related to the interaction with the environment. This has also facilitated further contact with related disciplines, such as neuroscience and ethology (Avila-García and Cañamero, 2002; Redgrave et al., 1999; Spier and McFarland, 1997). The cross-fertilisation of different scientific disciplines provides multiple benefits, e.g., a better understanding of the neural mechanisms underlying behaviour selection (Gurney et al., 2001b,a, 1998).

The animat approach was born in a concurrent and complementary fashion to these, proposing the study of developmental and evolutionary biology as sources of inspiration for modelling intelligent processes in a bottom-up fashion (Meyer, 1995; Wilson, 1991). In this context, the "animat" (the synthetic animal) is an ideal testbed of the aforementioned principles of intelligence, since it exhibits situatedness, embodiment and autonomy similarly to biological beings (Ziemke, 1998; Meyer, 1997). A concept which has gained my attention and partly overlaps with the aforementioned is *self-sufficiency*, since this provides a "baseline from which he (the experimenter) can judge the performance of the agent; either the agent is self-sufficient or not" (Spier and McFarland, 1997). The same authors also argue that the concept of self-sufficiency may provide a framework to assess the behavioural responses and the feedback from its own

internal state. The principles of learning proposed in previous chapters are examples of this feedback. Both the principles and the metrics have been inspired by ethological observations and measurements performed by experimental neuroscientists. Further discussion of this analogy will follow later in the text.

Assuming self-sufficiency as a criterion of optimality (Stephens and Krebs, 1986) is equivalent to providing an upper-boundary for performance, which depends on the particular agent studied. However, this equivalence is tautological, since the concept of optimality requires an observer to define it. Animals do not have this problem, since the criterion has emerged out of evolutionary processes and developmental interaction. The most elegant manner for solving the optimality issue for animats consists of applying the same evolutionary and developmental criteria experienced by animals. However, I have deliberately avoided this in the experimental approach. Instead of mimicking their processes at all levels, I have opted for a careful and reasonable analogy between animals and animats at a developmental level only. If animals have reached a good criterion to pair the feedback from the environment to their motivational state to be able to act appropriately, I can extract some analogous principles from their behaviour to use and test with our agent.

To frame this approach, the perspective of the animat focusing on the interaction with the environment as a source of intelligence, I draw on the ideas of J. J. Gibson in the field of ecological psychology (Gibson, 1966). He mostly focused on perception; however, he proposed the evolutionary nature of the cognitive processes (from perception to action selection) embodied by an animal. In fact, if I can assume that each animal and each species are the result of a process of mutual interaction with their niche, it should be straightforward to assume that their cognitive processes encompass the same principles. It was again the limitation of symbolic AI to provide an understanding of these cognitive processes which motivated a novel interest in the situated nature of cognition (Suchman, 1987) from the perspective of ecological psychology.

This thesis also adheres to the view of situated and embodied intelligence and to the necessity of providing not only feedback from the environment, but furthermore concrete criteria to analyse the underlying mechanisms of self-sufficiency. In this light, this study has focused on the principles giving rise to intelligence from an ecological perspective that grounds the agent to its environment: *affordance learning* and the *learning of behavioural patterns*.

Learning Affordances Situatedness has been one of the most often mentioned notions that an agent needs to exhibit intelligent behaviour (Ziemke, 1998). Beyond definitions, this dissertation has also addressed a study of the process of affordance learning as a possible way of situating an agent in its environment.

The approach I have followed in designing the architecture for learning affordances and the process itself has mainly followed three sets of ideas. Firstly, the ideas of Gibson on

ecological perception (Gibson, 1966). Gibson focused his attention on the interaction with the environment as a source of inspiration for understanding the visual system. The main conclusion of his work is that effective animal perception is shaped by the environment and by the needs of the animal. In a more formal manner, he defined the concept of affordance as the functional relationship between the perception of a cue or set of cues and a behaviour from the animal's repertoire. However, the notion of affordance has been heavily criticised due to the ambiguity with which Gibson proposed it and also partly due to the cognitivist mainstream in psychology. Secondly, behaviour based AI revisited ecological ideas in the early 80's establishing an initial conceptual framework to address intelligence from an ecological perspective. The third set of ideas follows as a natural consequence of this. AI scientists have looked at biology for inspiration to build synthetic agents. Coherently with ecology, perception has been viewed as one more process of interaction with the environment of the agent. In other words, from an ecological perspective it is not possible to consider perception, action selection and learning separately if the agent has to exhibit situatedness. Therefore, affordance learning is part of the process of situating the agent in its environment. Affordances will be related only to the agent itself, since these will not only depend on the objects or features of the environment, but also on every piece of the agent. The self-observation of the internal bodily dynamics and the relationship between these and the perception of and interaction with the environment is the principle driving the learning mechanism.

One of the contributions of this thesis has been a method to learn affordances in this framework. This contribution can also be viewed as a formalisation of a method to situate an agent. Indeed, the animat consists of a set of internal variables, drives, behaviours. By relating the fluctuations of the agent's internal dynamics to the execution of each behaviour and to the cues that were concurrently active when the behaviour was engaged, the agent is implicitly grounding the semantics of each interaction with regard to its internal needs and to its perception of the environment.

This learning method has been implemented taking into account two main relevant issues. Firstly that the amount of data provided by the sensors is very large. In this respect, it has been considered appropriate to apply data reduction techniques in the form of a Grow When Required (GWR) Network to dynamically group together sets of similar sensory patterns in a single topological node. On the one hand this simplification facilitates the analysis, on the other this network naturally adapts to newly perceived sensory patterns in a dynamic fashion. Secondly, a neural structure connecting the GWR nodal representation to the behaviour repertoire has been grown in a Hebbian manner, driven by the fluctuations of the internal physiology provoked by the execution of behaviours. Both the neural structure and the values of its synaptic weights are the affordances of the objects with respect to that agent. Deliberately, this neural structure is single layered, suggesting a direct relationship between the perception and the

action layers (a reactive architecture).

The adaptiveness of this architecture and of the learning process has been tested in an environment by varying the availability and the distribution of resources. The general conclusion is that for the given objects and situations it is possible to grow a neural structure to represent affordances in a dynamic fashion. The boundary to this is imposed by the causality between the perception of a sensory pattern and a certain fluctuation of its internal physiology. If both can be consistently related, the affordance of the behaviour involved in that fluctuation can be learnt. Both neural structures, the clustering (topological network) and the affordance networks, have been grown concurrently for biological coherence. However, there are still some controversial theoretical elements. First, the sensory system is based on snapshots of objects extracted from the sensory flow. This simplification has been assumed as a first approximation to reduce the amount of sensory information to be processed. Although this does not invalidate the experimental data, I suggest that the schema should be extended to handle sensory flow directly. I propose to do this by relating continuous time fluctuations of the agent's internal physiology to the Hebbian synaptic growing of the neural structure relating the topological nodes to the behaviours of the agent. This is an issue to be addressed in the future.

Furthermore, also from a *biological perspective*, I acknowledge that there is a set of internal processes, included in the process of affordance learning and excluded from the notion of grounding or situatedness. These could be grouped together under the label of affective phenomena; which modulate perception in an active manner. In fact, I do not perceive the function of the same object in the same manner when I am in different moods. This would probably introduce neuromodulation of perception, and the possibility of studying how animals adapt their perception according to an emotional state induced by itself and by its perception of the environment. Understanding these processes is a necessity to build new and more adaptive agents.

Biological Inspiration and Implication *Behaviour selection* has been individually addressed by different authors in the fields of animation (Blumberg, 1994), robotics (Ávila-García and Cañamero, 2004), and ethology (Spier and McFarland, 1997). However, the link that allows us to view behaviour selection as an adaptive process has arisen. Recent experiments and models in neuroscience (Fiorillo et al., 2003; Dayan and Balleine, 2002; Dayan, 2001; Suri and Schultz, 1998; Schultz, 1998; Schultz et al., 1997; Houk et al., 1995; Schultz et al., 1993) suggest that dopamine (DA) projections from the Ventral Tegmental Area (VTA) and Substantia Nigra Pars Compacta (SN_c) to striatal cells in the basal ganglia (BG) of some vertebrates acts as an error in the prediction of reward of a Stimulus-Response learning cycle (Pavlovian learning). Beyond the neuro-physiological evidence, striatal cells do receive synaptic projections from most parts of the cortex and from the VTA and SN_c, and project themselves to the BG output nuclei, the

Globus Pallidus Internal Segment (GP_i) and substantia nigra pars reticulata (SN_r). Furthermore, experiments show habituation of the DA signal when the cycle is repeated several times. Once the relationship is learnt, the error is zero.

However, this has been suggested for the case of Pavlovian contingencies only, hence when the release of reward is not mediated by the execution of any action or behaviour. In this thesis, I have extended this learning hypothesis for the case of instrumental learning, hence considering the cases for which reward is mediated by the execution of behaviour. This has direct consequences. Firstly, *behaviour selection* and *learning* are viewed as *two interacting processes* in a single framework. Secondly, this implicitly suggests that they partly share a common neural substrate and that one of the roles of Dopamine (DA) is to act as an error signal for learning behavioural patterns.

Motivation, Reward and the Actor Critic There have been two main sources of inspiration for embedding the actor-critic in the biologically inspired learning framework described throughout this thesis. These are Sutton and Barto (1981) and Houk et al. (1995). R. Sutton (in machine learning) suggested an algorithm for behaviour selection and learning in the framework of reinforcement learning: the actor-critic. This algorithm, unlike most reinforcement learning algorithms (e.g., Q-Learning (Watkins, 1989)), computes the error in the prediction and the policy to select behaviours separately. A critic relating states to the expected reward (SR fashion ones) calculates the error in the prediction of reward, and an actor learns policies to select behaviours to maximise the cumulative reward (Behaviour Selection). This same algorithm was also used by Houk (Houk et al., 1995) as a possible hypothesis for the processes embodied in the basal ganglia. Its separation between behaviour selection and learning makes it most attractive from a biological perspective. From the perspective of developmental biology, it seems reasonable to assume that simple animals do not need a critic, since their behaviour arbitration may work on a reactive basis. However, more complex animals have needed to modify their behavioural patterns on demand of the environment. The actor critic seems to be developed to this end. In this respect, evolutionary studies on the development of the biology of the basal ganglia may shed some light on this issue.

This view also relates to Bindra (1969), who suggested the presence of a *common substrate* for *reward* and *motivation*. This implicitly grounds the concept of self-sufficiency and optimality suggested by Spier and McFarland (1997) to the processes of interaction with the environment, based on reward and motivation. On the one hand, the motivational state is a combination of the internal physiological state of the agent (its internal drives) and of the state of the environment (its perceived affordances). On the other, the reward is mediated by the execution of a behaviour, a behaviour that has to be chosen in a manner that facilitates the agent's survival. At this point, it is convenient to draw on Ashby's notion of viability: "agents

need to be physiologically stable to survive" (Ashby, 1965). Following this, I have assumed that the feedback from the environment must be positive (good) when the execution of a behaviour leads towards values of higher stability and negative (bad) for other cases. This has also some theoretical implications, since I am assuming that the notions of reward and punishment correspond to positive and negative values, respectively.

From an ethological perspective, the environment and the agent's internal physiology interact with one another. The actor-critic has to lead this interaction towards behavioural patterns that maintain the viability of the agent. This has been tested in a series of scenarios, and the influence of the external and internal stimuli in forming these strategies has been quantified. This is further discussed in the next sections.

Neurological Implications Neuroscience has been one of the major sources of inspiration for this thesis. From this perspective, I would like to comment on some issues that the results introduced in the previous chapters have contributed to clarify. These are presented next.

- DA in the Basal Ganglia acts as the error in the prediction of reward following a certain stimulus. This view is supported by the experiments in the previous experiments, where the $\delta(t)$ (DA shot) is effectively signalling the error in the prediction of reward when this happens. This suggests that both the failure when selecting a behaviour is assessed by the DA signal, which makes learning possible. However, this must be dissociated from the notion of aversiveness, since learning in our case occurs after several iterations. The value of an aversive stimulus is learnt at the first trial. Therefore, I am suggesting that learning, via matching the right behaviour to maximise reward and failing at the selection, occurs in the same neural circuitry. Nevertheless, this does not directly enter the discussion on whether reward and punishment may be computed by different neural circuits, as suggested by Redgrave et al. (1999).

  However, it may also make sense that reward and punishment, once they have been calculated, enter the same neural circuit in order to modify the synaptic weights affecting the patterns of selection, since aversive stimuli tend to govern typical behavioural responses, e.g., to avoid the stimulus. Related to this, a recent hypothesis also argues that there might be different types of reward, whose compatibility with the hypothesis of the actor-critic would be interesting to investigate.

- In our architecture, the correction of the policies and of the critic itself is based on the difference between the expected reward calculated a priori for the execution of a behaviour and the real reward obtained a posteriori. This neuromodulatory behaviour is analogous to that assumed for DA in the basal ganglia. However, other authors argue that DA is only acting as a threshold that facilitates or complicates the release of a behaviour (Gurney et al., 2001b,a, 1998). In this respect I argue that both hypotheses may

be right. In fact, I fundamentally agree that the basal ganglia are acting as a centralised action selector. However, this is not its only role. Firstly, there is strong evidence for a learning role (Reynolds et al., 2001). Secondly, Redgrave et al. (1999) have argued that DA cannot be mediating the error in the prediction of reward since it cannot account for aversive stimuli. On the contrary, if, as I have assumed, punishment can be taken as negative reward, the hypothesis suggested by Schultz and Houk would hold. However, there is still no experimental evidence supporting this. Thirdly, if I assume that novel events are rewarding by themselves, it would be coherent that dopamine would signal events of this sort, as observed experimentally (Schultz et al., 1997).

Influence from The Environment The actor-critic has been hypothesised to be the embodiment of the biological algorithm modifying the agent's behavioural patterns. This has been tested in chapter 5 in a set of different environments. These have been engineered to cover a variety of situations, ranging from the scarcity of resources to their abundance. The actor-critic has been demonstrated to learn faster and to reach better values of internal physiological stability for the abundant environment than for the scarce. However, it has also been able to modify its behavioural patterns to adapt to situations where certain resources are scarcer than others by exhibiting reactive behaviour to that stimulus when this is presented. Unlike this, when an object affords more than one behaviour, our simulated agent tends to select the resource whose availability is lower. Therefore, the experimental results have demonstrated that the actor-critic is capable of modulating the agent's behavioural patterns, responding to variations in the availability and distribution of resources with appropriate behavioural cycles that lead the agent's deficits to zero in a successful manner.

Appetitive and Consummatory Behaviours Each behaviour considered throughout this thesis is motivation driven (Toates and Jensen, 1990). A motivation results from the combination of the the agent's internal drives and of the external affordances. In the architecture introduced in this thesis, motivations bias the agent's behavioural patterns on the basis of their intensity, which is in turn controlled by the expectation of reward. The larger the reward for a certain behaviour, the more likely that behaviour will be chosen for that motivational state in the future. This highlights that reward and motivation are intrinsically related, even for behaviours whose immediate reward is considered to be zero. These are the so-called *appetitive behaviours*, the ingraining of which in the competition for behaviour selection is still a controversial issue.

This thesis does not address this issue in its whole spectrum. However, the actor-critic architecture at hand has provided the opportunity of studying the case of a single appetitive behaviour (to avoid), which has been successfully integrated in the motivation driven architecture in an equal fashion to the remaining consummatory behaviours, see section 5.5.1. The experiments have shown that the behaviour "to avoid" is consistently executed when the agent is

sated and when the object at hand does not offer the possibility of satisfying any of the agent's drives. The agent therefore learns that it is more convenient to avoid spending any energy on a consummatory behaviour whose execution is not going to provide any reward and instead to execute the appetitive one which may lead at the next step to a positive reward.

This result is however bound by the assumption that the execution of the appetitive behaviour is less energy consuming than any consummatory one. This makes sense for the case of avoiding, since it can be argued that any consummatory behaviour is more energy consuming than this. However, it is important to stress this result, which is only valid for the behaviour to avoid, could also be extended to other appetitive behaviours with relatively simple modifications. So far the selection of the appetitive behaviour is possible because the policy that biases the selection has learnt to exclude the consummatory behaviours from that state. Therefore, the selection of the appetitive behaviour is performed *by exclusion* of the others. This could however change if the selection mechanism did not consist of selecting the winner policy (refer to section 5.3.2). This policy has served to faithfully explain a series of experimental cases. However, it can be argued that for complex organisms, exhibiting complex behaviours, an immediate benefit is not necessarily selected. This opens the possibility of theoretically learning to sequentially select appetitive behaviours when convenient for the given motivational state. In fact, if the policy update were not TD, it would be theoretically possible to integrate other appetitive behaviours in equal conditions to the consummatory ones in the competition. This is, however, to be addressed in the future.

To be consistent with the hypothesis that each behaviour is motivation driven, I have designed our decision framework to solely work on the basis of future reward, the same reward that alters the internal physiology and modifies the motivations themselves in preparation for the next selection. In conclusion, it is argued that appetitive behaviours are also motivation driven, and that they are used in the chain of events as a *filling* for the time gaps between two consummatory behaviours to maximise the global reward, therefore to maximise the physiological stability.

Internal Modulation I have also performed experiments to quantify the effect of the dynamics of the agent's internal physiology on the resulting behavioural patterns. Several different agents, each with a different physiology, were tested in scarce and abundant environments. The results demonstrate that the actor-critic can also explain influences other than external stimuli in the learning of behavioural patterns. If a homeostatic variable exhibits a fast decay, the agent will more likely select the behaviour that compensates this drive even if its value is not at a critical level; furthermore even if the resource needed to execute that behaviour is not very abundant in that scenario. On the contrary, if the homeostatic variable decays very slowly, any associated behaviour will be executed at most once or twice per cycle.

However, the experiments performed in chapter 6 have shown that the influence of the internal dynamics on the learning and on the behavioural patterns strongly depends on the environment. By comparing the behavioural patterns obtained for an abundant and for a scarce environment I can observe that in the former case the patterns turn out to be mostly driven by the agent's internal physiology. Conversely, for the latter case the agent mostly reacts to the affordance offered by the object encountered, hence exhibits reactive behaviour. However, the results show that in this case there is also some influence of the agent's internal physiology since for slow varying metabolisms the behavioural responses differ approx. 50% from the reactive pattern. The explanation for this is that for the case of behaviours exhibiting very slow internal physiological values, failing the execution of several behaviours within a cycle is insignificant in terms of reward and in terms of physiological stability. This highlights the idea that the emergence of the behavioural patterns depends on the combination of the internal physiology and of the external affordances. These cannot be considered as independent factors when analysing the behavioural patterns delivered by the actor-critic.

Another axis of influence of the agent's internal physiology is the function that relates effect to reward in the learning process. As mentioned at the beginning of this discussion, adaptation is driven in our agent by the reward resulting from the execution of behaviours. However, the procedure to calculate the reward related to a physiological effect due to the execution of a behaviour is controversial. To this end I have extended the definition of reward introduced by Schultz (Schultz et al., 1997) and have bound it to Ashby's notion of viability (Ashby, 1965). The principle consists of assuming that effects contributing to the agent's internal physiology are beneficial, or they are harmful. On this basis, the reward function introduced in chapter 5 delivers a positive value for the former case and a negative value for the latter. In this way I am extending the definition of reward to include as well the notion of punishment (negative reward) in a single framework. However, the biological plausibility of this extension seems still incomplete. A negative effect seems to be accounted for by different neural circuits depending on its intensity. If the effect is mildly negative the effect may be considered via the same method I have proposed. However, aversive experiences do not need to be repeated, they are learnt at once, hence suggesting that there are parallel neural pathways to learn about mildly negative and aversive situations. Much of this will have to be tested in future studies. However, a basic set of simulations studying the influence of the relationship between effect and reward has been illustrated in figure 5.2. The experiments have tested the sensitivity of varying the amount of effect in the same environments for the same behaviour execution, suggesting that a high amount of compensation is preferred for the given environment. This is reasonable, since this contributes to a higher internal physiological stability.

To conclude the discussion, the next subsection discusses the reach of the ecological principle in the approach I have followed throughout this thesis.

So far the Ecological Approach The principle of ecology has been mentioned on several occasions throughout this thesis as the criterion which drives adaptation and our design considerations in the animal world in an analogous fashion. This is the biological inspiration introduced and discussed at the beginning of this discussion chapter.

For *perception*, the ecological principle has helped to formalise a set of sensory patterns to be processed into a functional signal related to a set of behaviours within the agent's repertoire. Therefore, perception is a process intrinsically related to the way in which the agent elicits actions and interacts with its environment. Affordances imply an integrated view of the perception and action suggesting that an agent endowed with the capability of exploring its environment may be able to adapt to a variety of environments via interaction. Therefore this is a core idea of an attractive synthetic framework. Related to this, the model I have introduced is the first implementation of an affordance learning model embedded in a robotic platform. This has been tested in a variety of appositely engineered scenarios with different distributions of affordances. The learning is provoked by the interaction with the environment, which occurs when a behaviour is selected and executed. The resulting fluctuation provoked at the internal physiological level is then the drive modifying the neural substrate that represents the affordances and the policy to select behaviours in a way that promotes beneficial decisions in terms of physiological stability. The results have demonstrated that the agent has adapted to these environments by learning the affordances offered. By using this information, the agent can predict whether it will be able to perform that behaviour with that particular object.

Secondly, ecology has also been the inspiration for using the actor-critic in this schema. If affordances relate perception to the potentiality of performing an action, this part of the architecture drives and assesses the use of this information when *interacting with the environment*. The actor-critic in the animal world was hypothesised by Houk as a result of the observation of physiological and anatomical data in vertebrates as the algorithm that integrates learning and behaviour selection. However, besides the algorithm, a framework analogous to its natural context is required when used in a synthetic perspective. The neural substrate of the actor-critic in a vertebrate receives projections from every part of the cortex and the thalamus therefore suggesting that every piece of information, internal and external is assessed in this part of the brain for action selection. In an analogous fashion, our agent perceives data from its environment and from the internal physiology and learns to associate potentialities of action to every object and to every internal state. The consequences of this are that the affordances of the objects are filtered depending on the agent's motivational state giving rise to a set of policies to drive the agent's behaviour. It is important to stress that this is a way for the agent to ground knowledge with regard to its own perception of the environment.

Every learning process requires some assessment, which for the case of the actor-critic is the difference between the predicted and the real reward due to interaction with the environ-

ment. In a way, this closes the loop from perception to action in an ecological manner and satisfies the constraints imposed by ethological and neurological data.

In a way, this agent embodies a model of a learning architecture capable of adapting to a variety of environments provided that the agent has been endowed with the ability to explore. I have based our design principles on the idea that animals strive for survival and that the internal criterion to control this is the internal physiological stability. This is an interpretation of Ashby's notion of viability who suggested a set of essential physiological variables whose boundaries must be respected in order for the animal to exist. Furthermore, the architecture works in a biologically inspired manner. However, this architecture also exhibits some limitations both in terms of adaptation and of biological resemblance. These are addressed in the next and final section.

A Comparison with Redgrave's Model on Action Selection It has been considered appropriate to include a comparison between the model introduced by Redgrave, Gurney and Prescott (Redgrave et al., 1999; Gurney et al., 2001a) with the model introduced in this thesis at the end of this chapter.

Redgrave's model is based on the assumption that the main role of a vertebrate's basal ganglia is to arbitrate among the animal's behaviours in a centralised manner. The model introduced in this thesis also argues in favour of this hypothesis. However, our model also assumes other roles than selecting behaviours for the basal ganglia, such as learning behavioural patterns.

The robotic model of Redgrave, Gurney and Prescott is based on the assimilation of the biological hypothesis of action selection postulated by Redgrave: the basal ganglia is distributed into two subsystems exerting two complementary roles to select actions: selection and control. The cortex and the thalamus are considered to be *input* to the system, whose cells project to striatal cells and to the Sub-Thalamic Nucleus (STN). These structures project to the output nuclei, the Globus Pallidus Internal and External Segment (GP_i and GP_e, respectively). Furthermore, the projections from the Thalamus and the Cortices onto striatal cells are viewed as a set of parallel projections of sensory data. Based on these, the striatal cells elaborate the level of *saliency*, to contribute to disinhibit one output pathway or another in the GP_i. In order to do this calculation, striatal cells are supposedly organised in local recurrent networks, whose mutual interaction results in a saliency level for each pathway. As the reader may have already guessed, the incoming projections to the striatum are associated to the sensory input, while the projections from the GP_i back to the cortex and to the thalamus are considered to be the behavioural output in an analogous fashion to an artificial agent. Depending on the striatal cell receptor type, D1 or D2, the cell will inhibit or disinhibit, respectively, the nucleus it projects to. Therefore, D1 type cells, together with the projections of the STN, perform the selection

of behaviours per se, while the role of the D2 type cells is to act as a feedback loop increasing the saliency of those nuclei on the GP_i in such a manner that nuclei already active increase their activity, therefore magnifying the gradient of activity between the winner nucleus and its neighbours. Hence, if the nuclei in charge of the selection per se compose the selection pathway, those in charge of this magnification are labelled as the control pathway. This model suggest a very interesting process, both for the calculation of the saliency (which is computed in a single currency for each behavioural pathway), and for providing a reasonable explanation to the process of disinhibition itself.

Our model also makes similar assumptions with respect to the input and output projections as sensory and behavioural pathways, respectively. However, its inspiration, goals and reach are different. The model presented in this thesis is aimed at learning to perceive the environment and to learn behavioural patterns in a way that provides adaptation to the agent to live in this environment. Therefore, a comparison with the previous model is only partially possible. The anatomical elements shared by both models relate the cortical and thalamic projections to striatal cells, and these to the GP_i as the only pathway to select behaviours. However, I have also included the necessary anatomical elements to support the hypothesis that striatal cells are used to modify the calculation of the saliency for one or another behaviour depending on the relationship that striatal cells perceive from the cortex and thalamus via the dopamine projection from the Ventral Tegmental Area (VTA) and Substantia Nigra Pars Compacta (SN_c). Furthermore, our model does not make a distinction between striatal cells with D1 and D2 receptors. Instead, I focus our attention on the overall effect that dopamine has as a neuromodulator driving learning in these neural structures. This model supports the hypothesis that DA projections affect the synapses of these cells in terms of reward.

From an *experimental perspective*, both models share the aim of selecting roles, but there is a fundamental difference between them. Our model is suggesting the selection of motivation-driven actions based on the prediction of future reward that every behaviour may deliver. According to the hierarchy of adaptive processes introduced in chapter 2, our model implements behaviour selection, as does Redgrave's model; however, it also implements the learning of behavioural patterns in a biologically plausible manner. Therefore, it goes one level higher in the hierarchy of adaptive processes than Redgrave's model. By doing this, I have intrinsically related the concepts of motivation, reinforcement and reward, integrated in the same architecture to allow the modification of behavioural patterns. This argument has been extrapolated from the neurological and psychological literature. Furthermore support comes from the experiments of Schultz (Schultz et al., 1997), where it is suggested that it may also make sense that both reward and punishment, once they have been calculated, enter the same neural circuit in order to modify the synaptic weights affecting the patterns of selection, since aversive stimuli tend to govern typical behavioural responses, e.g., to avoid the stimulus.

In our architecture, the correction of the policies and of the critic itself is based on the difference between the expected reward calculated a priori for the execution of a behaviour and the real reward obtained a posteriori. This neuromodulatory behaviour is analogous to that assumed for DA in the basal ganglia. However, other authors argue that DA is only acting as a threshold that facilitates or complicates the release of a behaviour (Gurney et al., 2001b,a, 1998). In this respect I argue that both hypotheses may be right. In fact, I fundamentally agree that the basal ganglia are acting as a centralised action selector. However, this is not its only role. Firstly, there is strong evidence for a learning role (Reynolds et al., 2001). Secondly, Redgrave et al. (1999) have argued that DA cannot be mediating the error in the prediction of reward since it cannot account aversive stimuli. On the contrary, if, as I have assumed punishment can be taken as negative reward, the hypothesis suggested by Schultz and Houk would hold. However, there is still no experimental evidence supporting this. Thirdly, if I assume that novel events are rewarding by themselves, it would be coherent that dopamine would signal events of this sort, as observed experimentally (Schultz et al., 1997).

As for the model of Redgrave, an agent endowed with the same model has been capable of selecting behaviours and of doing so in a biologically plausible manner. Furthermore, it has been also possible for the same agent to modify its own behavioural patterns to adapt to changes in the environment based on the ecological principle that relates the environment to the agent's internal physiology. This has been demonstrated in a variety of cases in chapters 5 and 6. This suggests that this model may be correct in assuming that adaptation, lead by reward and punishment, may be working according to the principles listed throughout this thesis.

Future Work In this thesis affordances and internal physiology have been combined into a new learning framework, which enables adaptation by modifying the agent's own internal patterns and interacting with the environment.

However, ecology is a principle extending to dimensions beyond high level perception or behaviour selection. The model I have embedded in an agent adapts to the environment by exploiting the feedback provided by the environment in terms of reward in such a way that it modifies the neural weights encoding the agent's perception and behavioural patterns. Nevertheless, when comparing this model with a real animal interacting with its environment it becomes obvious that there are many interaction elements that continuously modify its structure in a manner that has not reproduced. Apart from highlighting the limitations of my model, this also spots a path to continue building ecological robots and agents. If this first model has mirrored the phenomenological processes playing a role in interaction with the environment, there are several ways to resume this approach:

- In the biological case, assessing the execution of a behaviour involves reward as modelled in our simulation environment, hence a signal of goodness or badness about the

action just performed. However, it is important to point out that the perception of reward is far more complex than a single signal reporting a feeling of satisfaction or disgust after the execution of a behaviour. Reward from a wider perspective involves other sorts of feedback than the post-proprioceptive feedback which has been used to calculate a simulated perception of reward for our agent. The feedback from the environment in its more general case precedes, encompasses and follows the execution of a behaviour involving a variety of phenomena. For example, proprioceptive and kinaesthetic feedback report information based on the self-perception of the execution of a behaviour and the resulting measure of reward directly depends on whether this feedback is correctly processed and used. Therefore, if future agents have to demonstrate adaptation from an ecological manner in complex environments and demanding tasks, I need to propose procedures to integrate these sorts of feedbacks in an ecological manner.

- From a phenomenological perspective, I have also modelled the learning of object affordances. Proprioception has been used in this case to reinforce or weaken the neural substrate that learns the affordances of a set of objects in the agent's environment. As extensions to this, the same suggestions introduced in the former point can be used to control the necessary mechanisms of neural plasticity to learn affordances.

- Related to this, I also have to argue that it is likely that the combination of several sources of feedback with the continuous processing of the sensory flow would be necessary to scale the model and the principles tested throughout this thesis.

- It is also necessary to point out that I have considered affordances in a very general manner. Affordances are the knowledge representing the potentiality of performing a behaviour when a set of cues is perceived. I have addressed this point by considering elementary object features or raw sensory data; other authors also argue that affordances are based on the integration of simple features that are separately extracted in the visual cortex of superior mammals (Tao et al., 2004), e.g., vertical and horizontal lines. I do not argue that this is a healthy way to continue in its most strict sense. However, recent studies of the visual cortex in macaques reveal that cells sensitive to horizontal cells and those sensitive to vertical cells are annealed in a complex non-linear recurrent network. This would suggest that these cells organise themselves after the ecological principle to interact with one another in such a manner that the appropriate information can be delivered to the F5 area in the pre-motor cortex to help to select behaviours. If these structures have emerged out of evolutionary and developmental processes to this end, it would be extremely interesting to understand the principles underlying the organisation of these neural structures from the perspective of an artificial agent.

- It is important to highlight that the model introduced by this thesis implements behaviour

selection as a process within a hierarchy of processes. According to this, learning and behaviour selection are processes that mutually influence one another. Therefore, it seems reasonable that the actor-critic were part of the implementation of these principles in the vertebrate's brain. However, this view shares and opposes that of Gurney and Prescott on how behaviour selection occurs. On the one hand the role of dopamine (DA) for Gurney and Prescott is viewed as a threshold signal to inhibit or facilitate the selection of behaviours. On the other dopamine acts for us as the error in the prediction of reward preceding the selection of a behaviour. A future issue to be addressed is the integration of both views on adaptation for behaviour selection and learning.

- Finally, it is also important to highlight that there are some limitations regarding the implementation. The estimation of the behaviour intensities and of the state value function of the critic has been performed via back-propagation of error (a variety of gradient descent). This has demonstrated to suffice for the environments at hand, therefore that this is a correct implementation from a phenomenological perspective. However, it would not be correct to assume that this is an intrinsically correct biological implementation of the phenomenon. Therefore, if a thorough biological model is to be designed, it would be necessary to transfer our perspective to the electro-physiological and biochemical level of description.

These possible extensions are considered from an ecological perspective but with an eye on the robotic applications and real problems that would need to be addressed in order to get a general model that integrates all the aforementioned processes. Unless it is proven wrong, the purpose of a model is to account for a set of phenomena it can explain. Hopefully the fate of this model is to become a particular case of a more general theory of ecological adaptation, where a single dynamics between the agent and its environment is thoroughly explained.

Chapter 8

Conclusion

This dissertation is about learning as a mechanism to adapt to a given scenario. In this respect, learning has been introduced from a biological perspective as a reward-driven process which guides perception and behaviour selection as interacting processes. Both are related via the ecological principle, which grounds the agent's perception and behavioural responses on the dynamics of its internal resources and on the dynamics of the interaction with the environment. Both these issues, ecological perception and the learning of behavioural patterns, have been addressed as phenomena ingrained in the adaptive process. To this end, a robotics architecture inspired on ecological and biological principles has been designed to test these hypotheses in a variety of meaningful situations.

The *perception* of this architecture is affordance-based. An affordance has been defined as the relationship between a cue or set of cues and a single behaviour, meaning the potentiality for a particular agent to perform that behaviour. This has straightforward implications in adaptive terms, since affordances implicitly represent the necessary knowledge to exhibit reactive behavioural patterns. Hence, from the perspective of an architecture, affordances could be seen as the lowest level of perception-action coupling in a hierarchical architecture for behaviour selection.

Behaviour selection and learning have been considered as two continuous levels of a hierarchy of adaptive processes. Behaviour selection from a traditional, ethological viewpoint has been formulated in an analytical fashion relating the agent's motivational state to the intensity of each behaviour. The selection consists then of choosing the behaviour exhibiting the highest intensity. The architecture introduced here fundamentally differs from this perspective, since a contribution of this thesis is the formulation of the selection of behaviour as a sub-problem of the learning of behavioural patterns. Based on the fact that learning by reinforcement can modulate the selection of behaviours, I have not imposed any specific formula to combine stimuli, since this is not required. Instead, behavioural patterns arise via interaction with the environment and are assessed via the reward signal. The actor-critic algorithm modifies the behavioural

patterns in such a manner that reward is maximised within a cycle of execution (leading the agent's deficits to their minimal values). The only constraint imposed has consisted of assigning a positive reward to effects of behaviour executions that contribute to diminish the agent's deficits and a negative value to those effects contributing in the opposite direction. These constraints have been suggested in the light of recent experiments in neuroscience and ethological observations.

The learning of object affordances has been addressed in chapter 4. This introduced the agent's model of internal physiology and a description of the internal homeostatic processes and external interactions regulating the learning and selection of behaviours. The learning of affordances is implicitly related to the effect that interacting (executing a behaviour) with an object may have. Based on this, two parallel processes, clustering (hence building a topological map) and synaptic growth (defining the weights of the synapses connecting each node to every behaviour) have been proposed and described as adaptation processes related to perception in analogy to those exhibited by animals. Two different topological network algorithms, GNG and GWR, have been tested in a variety of environments (varying in terms of shape and size of the objects, and of availability and distribution of objects). The results demonstrate that the process of clustering requires a minimum level of accuracy (represented by the final number of nodes of the topological map). Otherwise, the level of accuracy of the objects represented is not sufficient to grow functionally meaningful synaptic weights between the nodes and the behaviour repertoire of the agent. It is important to stress that the final synaptic weights arise through statistical (Hebbian) accounting of the effects that each individual behaviour provokes on the motivational state when executed in combination with every sort of object in the environment.

Chapter 5 addressed adaptation by considering behaviour selection and learning as two interacting processes. It firstly introduces the hierarchy of adaptive processes and the criterion that relates effect to reward. The models introduced address high-level behaviour selection in coherence with the hypothesis of dopamine (DA) mediating instrumental learning in the Basal Ganglia (BG). This suggests that this part of the brain also participates in the arbitration and engagement of goal directed actions. Therefore, low level, reflex movements or reward-unrelated actions would be unrelated to the conclusions reached within this context. The experiments have addressed the integration of external and internal stimuli for behaviour selection from an ecological perspective, varying in a set of appositely engineered scenarios, and in the availability and distribution of resources according to a set of hypotheses. A first and second version of the same architecture have been built and tested.

- The *first architecture* only considers the internal drives to bias one behaviour over another, with the final decision being computed via a multiplicative formula (intensity of each behaviour $\times$ perceived affordance values). For the proposed scenarios and for the

first proposed ethological boundary —maintain the stability of the *internal milieu*—, this is sufficient for scenarios with abundant resources. Nevertheless, the policies exhibited by agents using this architecture show a high level of opportunism, which has only demonstrated itself to be an intelligent strategy in scenarios where some resources are scarce.

- The *second architecture* introduced designs policies integrating the external and internal information as part of the state of the actor-critic algorithm. Therefore, the resulting decision patterns depend on the interaction with the environment, on the dynamic imposed by the learning algorithm and on the internal perception of reward. The experiments conducted have addressed these three degrees of freedom, suggesting the possibility of using the actor-critic in a wider range of scenarios.

The behavioural patterns obtained for the former case are solely learnt when the optimal behavioural strategy in terms of reward is to react to the multiplicative combination of affordances and drives. This is due to the decision making being solely based on the internal drives and on the stimulus closest to the agent, hence disregarding a global model of the environment. This limitation suggests that including the available affordances in the state of the actor-critic (as part of the motivational state) is fundamental for producing more accurate decisions. Given this architecture, the level of influence of external affordances and of internal stimuli has been considered to analyse the coherence of this behavioural architecture from an ethological and biological perspective. Interestingly, the results have suggested that this architecture is capable of generating appropriate behavioural patterns to maintain the internal deficits within the agent's viability zone. Furthermore, it has also showed that, if appetitive and consummatory behaviours are competing in a motivation driven schema, the actor-critic will be directly involved in calculating the prediction of reward of the execution of each behaviour (appetitive and consummatory), competing in a similar fashion for the agent's actuators.

Finally, despite the set of successful contributions, it is necessary to stress that the integration of learning and behaviour selection in the aforementioned actor-critic is insufficient to explain several ethological phenomena. Ethology shows examples of behaviour selection not based on equal evaluation of each behaviour, but according to groups or habits. The architectures in this thesis are not sufficient to explain these. Furthermore, acquired knowledge relies on the weights of the neural structures predicting reward in the actor-critic. This is a plastic structure, which continues representing the acquired knowledge if there is a continuous exposition to the circumstances on the environment. This does not always happen in the animal realm.

The model introduced suggests the existence of parallel learning and distributed decision making processes distributed at the different levels of the hierarchy. The same reinforcement

learning process used for learning can also be extended to learn policies for behaviour selection as a process of habituation depending on the appropriateness of the stimuli and on the behaviour's frequency of execution. Hence, under normal circumstances, the execution of one or another behaviour will be highly dependent on the stimulus.

It is also important to remark that the level of behaviour selection addressed by the basal ganglia does involve high-level behaviours (not low level actions, at least directly) and only reward-directed actions (not reactive actions). This is supported by clinical evidence, since Parkinson's patients (whose release of dopamine is fairly reduced) do not perform motivated actions. However, they do react to stimuli in an unconditioned fashion.

Appendix A

Hypothesis of Correspondence between Webots and the New Simulator

The main issue addressed in chapter 5 is *the learning of policies* in a strategy for behaviour selection with the requirement that physiological stability is maintained. The experimental testbed currently consists of a Khepera robot, simulated with Webots 4.0. The use of this simulator imposes severe restrictions from a practical viewpoint, since the performance of experiments is bounded by the availability of Webots licenses and by the length of the simulations. Each simulation extends over 10 hours, which motivates the use of an optimised implementation for the problems addressed in this chapter.

The most time-consuming element of the simulator is the graphics-engine, which has been by-passed by building a simulator where the environment and the interaction have been abstracted. The interactions have been simulated on the basis of a formula that calculates the affordance of an object as a function of the features of the object. Each object has a set of affordances depending on its physical features. It has been assumed that the agent has this knowledge. This simplification enables a dramatic reduction of the simulation length, while still enabling the cognitive issues on the combination of stimuli and behaviour selection to be addressed with a sufficient level of correspondence with the real interaction.

This section addresses *the level of correspondence* in a dual fashion. Firstly, it shows the mathematical similarities between the two simulators. Secondly, it introduces a set of experiments to demonstrate that at a cognitive level, there exists an appropriate level of correspondence between the simulated and the real environment. To that aim, the affordance values used for behaviour selection have been intentionally distorted via the addition of white noise in order to evaluate the degradation of the learning and selection process. Results obtained with the Webots simulator and the new simulator are compared.

A.1 Mathematical Framework

The elements of the architecture introduced in section 5.3.2 are valid for both simulation environments. The learning is due to an actor-critic (AC) reinforcement learning architecture, a variant of a Temporal Difference (TD) algorithm, where the assessment of decisions and the execution of behaviour are embodied in two separate modules: the critic and the actor, respectively. The learning operation is illustrated by equations 5.5, 5.6, 5.3 and 5.4.

For the Webots simulator, the robot navigates among the objects in a random fashion. Everytime it encounters an object, a decision is made and a behaviour executed, whose effect reflects on the internal physiology. Unlike this, for the new simulator, there is no navigation. The "encounter" with an object is reduced to the random selection of one of the objects to interact with.

There is randomness for both cases. Nevertheless, there is a fundamental difference. The navigation for the case of the Webots simulator is forcing some patterns of encounter of the different objects in the environment due to their physical distribution. These patterns are absent for the case of the abstract environment, since the selection of objects obeys a random policy. In this respect, equation A.1 expresses the general case of a transition probability.

$$Pr\{s_{t+1} = s', r_{t+1} = r | s_t, r_t, b_t, s_{t-1}, r_{t-1}, b_{t-1}, ...\} \tag{A.1}$$

In equation A.1 the future state s_{t+1} and future reward r_{t+1} depend on previous states (s_t, s_{t-1} ...) and on previous reward value (r_t, r_{t-1} ...) and behaviours executed (b_t, b_{t-1} ...).

However, the structure of the environment allows us to re-formulate this dependence of every state transition according to the Markovian property, where every state s_{t+1} depends on the precedent state s_t only. This difference has an effect on the structure of the space of states. In fact, the Markov property states that the probability of being in state $s_{t+1} = s'$ depends on the previous state only, see equation A.1. This means that the probability of moving from one state to the next depends on the initial state only. By learning the policies, we are adjusting the probabilities of moving from one initial state s_t to the goal, in state s_{t+k} so that the cumulative reward is maximised. At this point it is necessary to stress that the state is divided into two parts, the drives d_t and the affordances a_t. The former experience an exponential decay when not affected by an interaction, hence, their state is predictable. However, the affordances cannot be predicted because the object to interact with in the abstract environment is chosen at random. At this point, it becomes obvious that predicting the next state from the current state is impossible, since only the d_t part of the state is predictable, see equation A.2.

$$Pr\{s_{t+1} = s', r_{t+1} = r | d_t, a_t\} = Pr\{s_{t+1} = s', r_{t+1} = r | \{d_t, a_t\}, b_t\} \tag{A.2}$$

Nevertheless, it is still possible to learn on the basis of single transitions, since the reward

does not depend on the final affordance (a_{t+1}), but only on the affordances available at the moment of decision (a_t):

$$\vec{\omega}_{t+1} = \vec{\omega}_t + \alpha\delta_t \vec{e}_t, \tag{A.3}$$

where $\vec{\omega}_{t+1}$ is the vector of weights used to estimate the policy values for each behaviour and the prediction of future reward. e_t is the error vector that re-scales the scalar value of the error in the prediction of reward d_t for every weight.

The validity of these equations is guaranteed even in an environment where the encounter of objects is not predictable. However, the use of *eligibility traces* (TD(λ)) is, in these environments, excluded. Therefore, the learning is less efficient from the mathematical perspective, than it in the case of Webots because the environment cannot be foreseen. However, the abstraction of the interaction makes it possible to reduce the real time from hours to minutes, which facilitates the performance of experiments.

Eligibility traces are an implementation of heuristics applied to the TD algorithm. It is known that the solution is a trace of transitions from the origin to the goal state that maximise reward. Applying elibigility traces is equivalent to extending the event of getting reward from the current to the previous transition, see A.4 —the right term propagates the gradient of the cumulative reward from the current to the previous states. This creates the so-called elibigibility trace. Algorithmically, this has an effect on the update of the weights of the networks estimating the policies and the function of the critic, cf. equations A.3 and A.4.

$$\vec{e}_t = \gamma\lambda\vec{e}_{t-1} + \nabla_{\vec{\omega}_t} V_t(s_t) \tag{A.4}$$

The next section introduces a set of experiments to demonstrate that despite the differences introduced in this section, there is nevertheless a correspondence between the results obtained in one experimental testbed and the other. We have first performed the experiments with the Khepera in Webots 4.0, and then replicated the simulations with the same parameters in the new simulator.

A.2 Degrading Affordances for both Simulators: a Comparative Study

The goal of the first set of experiments is to study the level of dependence of the learning process on the affordances. To that aim, distortion has been gradually added to the initial affordance values ($a_i'(t)$) in the form of white additive noise ($n(t), m_x = 0, amplitude = \alpha$), as shown in equation A.5. Sets of 2 simulations have been run for each value of noise, varying its amplitude (α) between 0 and 1 in increments of 0.2. The length of each simulation has been 2×10^5 decisions (*ca.* 5×10^5 time steps, 7 hours per simulation).

$$a_i(t) = (a_i^{'}(t) - 0.5) * (1 - \alpha) + 0.5 + \alpha * n(t) \tag{A.5}$$

Equation A.5 shows the affordance value resulting from the addition of gaussian noise ($n(t), m_x = 0, amplitude = \alpha$) to its original value ($a_i^{'}(t)$).

The learning is organised in episodes. Each of them commences by setting the homeostatic variables to random values between 0.0 and 1.0. The agent will then have to make appropriate decisions until the norm of the vector of deficits (the drives) is smaller than 0.1 (the boundary of the optimal zone), cf. figure 5.2. A new episode then follows. This process repeats throughout 5×10^5 simulation steps.

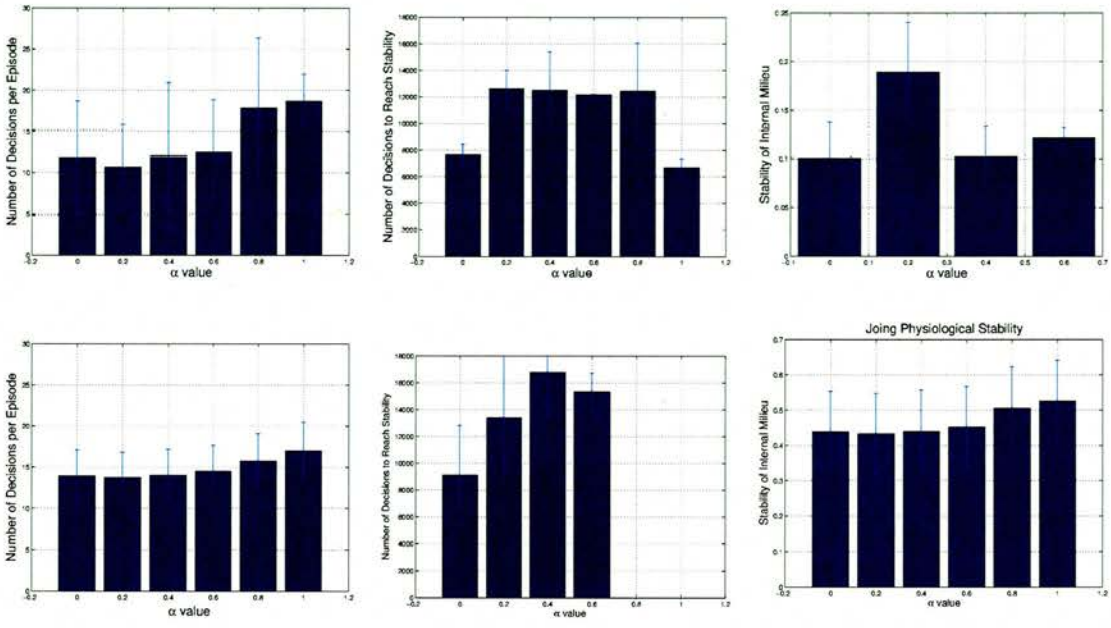


Figure A.1: From left to right: Mean Number of Decisions per Episode. Number of Decisions to Reach Stability. Physiological Balance and Overall Comfort, mean and variance values, respectively. Top graphs show results for the Khepera Robot in the Webots Simulator, bottom graphs show results for the new simulator.

Results with Webots Figure A.1 shows the mean number of decisions per episode (needed to reach the optimal zone of the physiological space) at the end of the simulation for each value of α. This suggests that affordances improve the performance by reducing the number of decisions needed to reach the viability area to about 12 decisions. Also, the performance decreases as the level of noise increases. For values of α larger than 0.6, the higher number of decisions required suggests that the affordance values are disregarded, turning decision-making into a process only controlled by the drives (hence disregarding the possible affordances offered by the object faced). Figure A.1 (centre graph) shows the number of episodes needed to reach

stable policies. The time needed to reach convergence when there is no distortion is 7000 decisions, dramatically increasing for higher levels of distortion. The value for $\alpha = 1.0$ signals that in this case, the policies never get to their right values, hence in this case only, it takes few time steps to reach a bad initial (and final) policy.

Results with the new Simulator The new experimental testbed *simulates the interaction.* The environment itself consists of a vector containing the objects in the environment, each with different affordance values. Navigation is absent in this environment, since the interaction consists of randomly selecting an object from the vector, to which the same interaction rules are applied as in the real scenario.

Graphs A.1 (bottom set) introduce the decrease in performance in terms of the number of decisions per episode to reach stability, tested by re-setting the internal physiology values to random values between 0 and 1, when convergence for the policies has been reached, and number of decisions to reach stability, left and centre, respectively. The bars are parametrised according to the distorting value α (cf. equation A.5).

Simulations have been run with for same decay rate value $\tau = 5x10^{-4}$ used for the simulations in the real scenario. Those to measure physiological stability have been run for both cases with $\tau = 5x10^{-3}$.

Figure A.1 suggests that affordances improve the performance by reducing the number of decisions needed to reach the viability area to approx. 12 decisions. Also, the performance decreases as the level increases.

Conclusion

Previous sub-sections have addressed the *the level of correspondence* between the simulations performed with the Webots simulator and the new simulator. Webots offers the advantage of being a realistic simulator, using physics to model the interaction with the objects. Nevertheless, in this section it has been argued that a faster simulation environment is able to provide corresponding results with respect to the cognitive issues addressed in the rest of this chapter.

To that aim, the mathematical properties for both simulation environments have been compared, highlighting that due to the random selection of objects in the new simulator, it is only possible to use the simplest version of the Temporal Difference algorithm (TD(0)). Furthermore, some experiments to measure the impact of a simulated distortion of the affordance weights on the learning process and the behaviour selection performance have been run with both simulators. The results are introduced in figure A.1.

- By comparing the *number of decisions to reach the goal* (left graphs), it can be concluded that with minor differences, the number of decisions experiences a similar degradation.

The best and worst values are, for the Webots simulator 11 and 18, for the simulated environment 13 and 17, respectively. The tendency to degrade when the level of noise increases (α) shows the same pattern, the more noise, the more difficult it becomes to reach the optimal zone.

- The middle graphs represent the *number of decisions needed to learn the policy* (that maintain stability). A similar pattern of degradation, proportional to the level of noise, is shown in both graphs. The best and worst values are 8800 and 12500 for Webots, and 9000 and 18500 for the simulated world. Although the pattern is similar, the numerical difference can be explained by the fact that the update happens on the basis of single, unpredictable transitions in the second case. This seems to be more sensitive to high levels of distortion. An exception to the general behaviour is the values obtained for α between 0.8 and 1.0. This is because the level of stability reached is not within the optimal zone. Not surprisingly, reaching a bad level of stability requires a shorter time.

- Finally, the right graphs show a comparison in terms of *physiological stability*. Again, it can be shown that the values obtained in both cases show a similar pattern of degradation proportional to α.

This set of experiments has addressed the learning of policies for behaviour selection using TD(0) with the proposed Actor-Critic architecture. The selected set of experiments demonstrate that although there are some differences for the interaction which affect the speed of convergence of the learning algorithm, there is a correspondence in terms of the cognitive aspects addressed in this chapter. In fact, the learning of behaviours obeys the same rules, and the influence of the sensory perception (the affordances) has a similar effect, as shown by the three metrics used for the comparison.

On the basis of this demonstrated correspondence, and motivated by the serious limitations imposed by the Webots simulator, it is proposed to conduct the experiments in chapters 5 and 6.

Bibliography

Ackley, D. and Littman, M. (1991). Interactions between learning and evolution. In Langton, C., Taylor, C., Farmer, J. D., and Rasmussen, S., editors, *Artificial Life II, SFI Studies in the Sciences of Complexity*, volume X, pages 487–509. Addison-Wesley.

Agre, P. and Chapman, D. (1987). Pengi: An implementation of a theory of activity. In *Proceedings of the Sixth National Conference on Artificial Intelligence*, pages 268–272.

Ashby, W. (1965). *Design for a Brain: The Origin of Adaptive Behaviour.* Chapman & Hall, London.

Avila-García, O. and Cañamero, L. (2002). A comparison of behaviour selection architectures using viability indicators. In *Proc. of International Workshop on Biologically-Inspired Robotics: The Legacy of W. Grey Walter.* Bristol HP Labs, UK.

Ávila-García, O. and Cañamero, L. (2004). Using hormonal feedback to modulate action selection in a competitive scenario. In Schaal, S., Ijspeert, A., Billard, A., Vijayakumar, S., Hallam, J., and Meyer, J.-A., editors, *In Proceedings of the 8th International Conference on Simulation of Adaptive Behavior (SAB'2004).* MIT Press.

Baecker, R., Grudin, J., Buxton, W. A., and Greenberg, S. (1995). *Readings in Human-Computer Interaction: Towards the Year 2000.* Morgan Kaufmann.

Baerends, G. (1976). The functional organisation of behaviour. *Animal Behaviour*, (24):726–735.

Bernard, C. (1878). *LeÇons sur les phénomènes de la vie communs aux animaux et aux végétaux.* Paris : J.-B. Baillière.

Billard, A. and Hayes, G. (1998). Drama, a connectionist architecture for control and learning in autonomous robots. *Adaptive Behavior.*

Billard, A. and Mataric, M. J. (2001). Learning human arm movements by imitation: Evaluation of a biologically inspired connectionist architecture. *Robotics and Autonomous Systems*, 37(2–3):145–160.

Bindra, D. (1969). The interrelated mechanisms of reinforcement and motivation, and the nature of their influence on response. In Arnold, W. J. and Levine, D., editors, *Nebraska Symposium on Motivation*, pages 1–33. University of Nebraska Press.

Blumberg, B. (1994). Action selection in hamsterdam: Lessons from ethology. In *Proceedings of Third Intl. Conference on Simulation of Adaptive Behaviour (SAB94)*. Cambridge MA: MIT Press.

Blumberg, B. (1997). *Old Tricks, New Dogs: Ethology and Interactive Creatures*. PhD thesis, Massachussets Institute of Technology.

Braitenberg, V. (1984). *Vehicles: Experiments in Synthetic Psychology*. MIT Press.

Brooks, R. (1986). A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, (RA-2):14–23.

Brooks, R. (1990). Elephants don't play chess. *Robotics and Autonomous Systems*, 6:3–15.

Bryson, J. J. (2000). The study of sequential and hierarchic al organisation of behaviour via artificial mechanisms of action selection. M. Phil. at the University of Edinburgh.

Bryson, J. J. (2004). Action selection and individuation in agent based modelling. In Sallach, D. L. and Macal, C., editors, *The Proceedings of Agent 2003: Challenges of Social Simulation*.

Cañamero, L. D. (1997). Modeling motivations and emotions as a basis for intelligent behavior. In Johnson, W. L., editor, *Proceedings of the First International Symposium on Autonomous Agents (Agents'97)*, pages 148–155. New York, NY: ACM Press.

Cannon, W. B. (1929). *The wisdom of the body*. London : Kegan Paul, Trench, Trubner & Co.

Chapman, D. (1987). Planning for conjunctive goals. *Artificial Intelligence*, 32(3):333–377.

Cooper, R. and Glasspool, D. (2002). Learning affordances and action schemas. In French, R. and Sougne, J., editors, *Connectionist Models of Learning, Development and Evolution*, pages 133–142. Springer-Verlag: London.

Cos-Aguilera, I., Cañamero, L., and Hayes, G. (2003). Motivation-driven learning of object affordances: First experiments using a simulated khepera robot. In Detje, F., Dörner, D., and Schaub, H., editors, *The Logic of Cognitive Systems. Proceedings of the Fifth International Conference on Cognitive Modelling, ICCM*, pages 57–62. Universitäts-Verlag Bamberg.

Crook, P. A. and Hayes, G. M. (2003). Learning in a state of confusion: Perceptual aliasing in grid world navigation. In *In Proceedings of TIMR 2003 - Towards Intelligent Mobile Robots*, UWE, Bristol. UWE.

Damasio, A. R. (1999,2000). *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. Harcourt Brace.

Damoulas, T. A. (2004). Evolving a sense of valency. Master's thesis, IPAB, School of Informatics, The University of Edinburgh.

Darwin, C. (1866). *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. John Murray, 4th edition.

Dawkins, R. (1976). Hierarchical organisation: A candidate principle for ethology. In Bateson, P. and Hinde, R., editors, *Growing Points in Ethology*. Cambridge University Pres.

Dayan, P. (2001). Motivated reinforcement learning. In *Proceedings of NIPS 2001*.

Dayan, P. and Balleine, B. W. (2002). Reward, motivation and reinforcement learning. *Neuron*, 36:285–298.

Demiris, Y. and Hayes, G. (2002). *Imitation as a dual-route process featuring predictive and learning components: a biologically-plausible computational model*, volume Imitation in Animals and Artifacts, chapter 13. MIT Press.

Doya, K. (1999). Reinforcement learning in continous time and space. *Neural Computation*.

Fagg, A. H. and Arbib, M. A. (1998). Modeling parietal–premotor interactions in primate control of grasping. *Neural Networks*, 11(7-8):1277–1303.

Fellous, J.-M. (2001). Dopamine modulation of prefrontal delay activity-reverberatory activity and sharpness of tuning curves. *Neurocomputing*, (38–40):1549–1556.

Fellous, J.-M. (2004). From human emotions to roboti emotions. In Cañamero, L. and Hudlika, E., editors, *Architectures for Modelling Emotion: Cross-Disciplinary Foundations. Papers from the 2004 AAAI Spring Symposium*. AAAI Press.

Fellous, J.-M. and Suri, R. E. (2003). *The Handbook of Brain Theory and Neural Networks*, chapter The Roles of Dopamine. MIT Press.

Fiorillo, C. D., Tobler, P. N., and Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, 299:1898–1902.

Fisher, R. A. (1930). *The genetical theory of natural selection*. Clarendon Press, Oxford.

Freud, S. (1940). *An Outline of Psychoanalysis*. Hogarth Press.

Fritzke, B. (1994). A growing neural gas network learns topologies. In Tesauro, G., Touretzky, D. S., and Leen, T., editors, *Advances in Neural Information Processing Systems*, number 7, pages 625–632. MIT Press.

Gadanho, S. (2002). Emotional and cognitive adaptation in real environments. In Trappl, R., editor, *Cybernetics and Systems 2002. Proceedings of the 16th European Meeting on Cybernetics and Systems Research. ACE'2002 Symposium.*, volume 2, pages 762–767. Austrian Society for Cybernetic Studies.

Gadanho, S. C. (1998). *Reinforcement Learning in Autonomous Robots: An Empirical Investigation of the Role of Emotions*. PhD thesis, University of Edinburgh.

Gärdenfors, P. (1992). *Belief Revision: An Introduction*. Cambridge University Press.

Gershenson, C. (2001). Artificial societies of intelligent agents. Master's thesis, National University of Mexico.

Gershenson García, C. (2000). Action selection properties in a software simulated agent. In et al., C., editor, *MICAI 2000: Advances in Artificial Intelligence*, number 1793 in Lecture Notes in Artificial Intelligence, pages 634–648. Springer Verlag.

Gibson, J. (1986). *The Ecological Approach to Visual Perception*. Hillsdale, N.J. ; London.

Gibson, J. J. (1966). *The Senses Considered as Perceptual Systems*. Houghton Mifflin Company, Boston.

González, P. P., Negrete, J., Barreiro, A., and Gershenson, C. (2000). Action selection properties in a software simulated agent. In et al., C., editor, *MICAI 2000: Advances in Artificial Intelligence*, number 1793 in Lecture Notes in Artificial Intelligence, pages 621–633. Springer Verlag.

González Perez, P. P., Negrete Martínez, J., Barreiro García, A., and Gershenson García, C. (2000). A model for combination of external and internal stimuli in the action selection of an autonomous agent. In *MICAI 2000: Advances in Artificial Intelligence. Lecture Notes in Artificial Intelligence*, number 1793 in Lecture Notes in Artificial Intelligence, pages 621–633. Springer-Verlag.

Guazzelli, A., Corbacho, F. J., M., B., and Arbib, M. (1998). Affordances, motivation, and the world graph theory. *Adaptive Behavior*, (6(3/4)):435–471.

Gurney, K., Prescott, T., and Redgrave, P. (2001a). A computational model of action selection in the basal ganglia. i. a new functional anatomy. *Biological Cybernetics*, 84:401–410.

Gurney, K., Prescott, T. J., and Redgrave, P. (1998). The basal ganglia viewed as an action selection device. In *Proceedings of the Eigth International Conference on Artificial Neural Networks*, pages 1033–1038.

Gurney, K., Prescott, T. J., and Redgrave, P. (2001b). A computational model of action selection in the basal ganglia. ii. analysis and simulation of behaviour. *Biological Cybernetics*, (84):411–423.

Hallam, J. C. T. (1998). Can we mix robotics and biology? In *Proceedings of IROS'98*.

Harnad, S. (1990). The symbol grounding problem. *Physica D*, 42:335–346.

Hebb, D. (1949). *The organization of behavior. A neuropsychological theory*. Wiley, New York.

Heyes, C. M. and Galef, B. G. J., editors (1996). *Social Learning in Animals: the roots of culture*. San Diego : Academic Press.

Hinde, R. A. (1960). Energy models of motivation. In *Symposium of Society of Experimental Biology*, volume 14, pages 199–213.

Hinde, R. A. (1971). Critique of energy models of motivation. In Bindra, D. and Stewart, J., editors, *Motivation*, pages 36–48. Penguin.

Houk, J. C., Adams, J. L., and Barto, A. G. (1995). Models of information processing in the basal ganglia. In Houk, J. C., Davis, J. L., and G., B. D., editors, *A Model of How the Basal Ganglia Generate and Use Neural Signals That Predict Reinforcement*, A Bradford Book, chapter 13, pages 249–270. MIT Press, 2nd. edition (1998) edition.

Hull, C. (1943). *Principles of Behaviour: an Introduction to Behaviour Theory*. D. Appleton-Century Company, Inc.

Humphries, M. (2002). *The basal ganglia and action selection: A computational study at multiple levels of description*. PhD thesis, University of Sheffield.

Humphrys, M. (1997). *Action Selection methods using Reinforcement Learning*. PhD thesis, Trinity Hall, University of Cambridge.

Izard, C. (1993). Four systems for emotion activation: Cognitive and noncognitive processes. *Psychological Review*, 100(1):68–90.

Kessen, W., Levine, J., and Wendrich, K. (1969). The imitation of pitch in infants, infant behavior and development. *Infant Behavior and Development*, (2):93–99.

Kohler, E., Keysers, C., Umiltà, M. A., Fogassi, L., Gallese, V., and Rizzolatti, G. (2002). Hearing sounds, understanding actions: Action representation in mirror neurons. *Science*, 297:847–848.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69.

Kohonen, T., Kaski, S., and Lappalainen, H. (1997). Self-organized formation of various invariant-feature filters in the adaptive-subspace som. *Neural Computation*, 9(6):1321–1344.

Konidaris, G. D. (2003). Behaviour based robotics. Master's thesis, IPAB, School of Informatics, The University of Edinburgh.

Kravitz, E. A. (1988). Hormonal control of behaviour: Amines and the biasing of behavioral output in lobsters. *Science*, (241):1175–1781.

Lin, L.-J. (1993). *Reinforcement Learning for Robots Using Neural Networks*. PhD thesis, Carnagie Mellon University. Technical Report CMU-CS-93-103.

Lorenz, K. (1966). *Evolution and Modification of Behaviour*. Methuen & Co Ltd, London.

Lorenz, K. (1971). *Studies in animal and human behaviour*, volume 2. Methuen. London.

Maes, P. (1991). A bottom-up mechanism for behaviour selection in an artificial creature. In Meyer, J. and Wilson, S., editors, *Proceedings of the First International Conference on Simulation of Adaptive Behavior*, pages 238–246. Cambridge MA: MIT Press.

Maes, P. (1997). Modeling adaptive autonomous agents. In Langton, C. G., editor, *Artificial Life. An Overview*, pages 135–166. The MIT Press.

Maistros, G. and Hayes, G. (2001). An imitation mechanism for goal-directed actions. In *Proceedings TIMR 01 - Towards Intelligent Mobile Robots, Manchester*, number UMCS-01-4-1 in Technical Report Series. Manchester University.

Marée, A. F. M., Panfilov, A. V., and Hogeweg, P. (1999). Phototaxis during the slug stage of *Dictyostelium discoideum*: a model study. *Proceedings of the Royal Society of London. Series B. Biological sciences.*, 266:1351–1360.

Marsland, S., Shapiro, J., and Nehmzow, U. (2002). A self-organising neural network that grows when required. *Neural Networks*, 15(8-9):1041–1058.

Mataric, M. J. (2000). Getting humanoids to move and imitate. *IEEE Intelligent Systems*, pages 18–24.

Maturana, H. R. and Varela, F. J. (1980). *Autopoiesis and Cognition: the Realization of the Living*, volume 42 of *Boston Studies in the Philosophy of Science*. Dordrecht; London: Reidel.

McClure, S. M., Daw, N. D., and Montague, P. R. (2003). A computational substrate for incentive salience. *TRENDS in Neurosciences*, 26(8).

McDougall, W. (1913). The sources and direction of psychophysical energy. *American Journal of Insanity*.

McFarland, D. (1990). What it means for robot behaviour to be adaptive. In Meyer, J.-A. and Wilson, S. W., editors, *Proceedings of the First International Conference on Simulation of Adaptive Behavior*, pages 22–28. MIT Press.

McFarland, D. (1993). *Animal Behaviour*. Prentice Hall, third edition edition.

McFarland, D. and Spier, E. (1997). Basic cycles, utility and opportunism in self-sufficient robots. *Robotics and Autonomous Systems*, (20):179–190.

McFarland, D. J. and Sibly, R. M. (1975). The behavioural final common path. *Proceedings of the Royal Society of London*, 270:265–293.

Metta, G. and Fitzpatrick, P. (2002). Better vision through manipulation. In Prince, C. G., Demiris, Y., Marom, Y., Kozima, H., and Balkenius, C., editors, *Proceedings Second International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems 94*, pages 97–104.

Meyer, J.-A. (1995). The animat approach to cognitive science. In Roitblat, H. L. and Meyer, J.-A., editors, *Comparative Approaches to Cognitive Science*, pages 27–44. MIT Press.

Meyer, J.-A. (1997). From natural to artificial life: Biomimetic mechanisms in animat designs. *Robotics and Autonomous Systems*, (22):3–21.

Nehaniv, C. and Dautenhahn, K. (1998). Mapping between dissimilar bodies: Affordances and the algebraic foundations of imitation. In Demiris, J. and Birk, A., editors, *Proceedings of the European Workshop on Learning Robots 1998 (EWLR-7), Edinburgh, Scotland.*

Noble, J. (1998). Cooperation, conflict and the evolution of communication. *Adaptive Behaviour*.

Oztop, E. and Arbib, M. A. (2002). Schema design and implementation of the grap-related mirror neuron system. *Biological Cybernetics*, (87):116–140.

Pavlov, I. P. (1927). *Conditioned reflexes : an investigation of the physiological activity of the cerebral cortex*. Oxford University Press.

Peters, J. (2003). Reinforcement learning for humanoid robots - policy gradients and beyond. In *Proceedings of the Third IEEE International Conference on Humanoid Robotics 2003*.

Pfeifer, R. (1994). The fungus eater approach to emotion: a view from artificial intelligence. *Cognitive Studies*, 1:42–57.

Prescott, T. J. (2001). The evolution of action selection. In Holland, O. and McFarland, D., editors, *The whole iguana*. Cambridge MA: MIT Press.

Redgrave, P., Prescott, T., and Gurney, K. (1999). The basal ganglia: a vertebrate solution to the selection problem? *Neuroscience*, 89:1009–1023.

Reynolds, J. N. J., Hyland, B. I., and Wickens, J. R. (2001). A cellular mechanism of reward-related learning. *Science*, (213):67–70.

Rizzolatti, G., Fogassi, L., and Gallese, V. (2000). Cortical mechanisms subserving object grasping and action recognition: A new view on the cortical motor functions. In Gazzaniga, M., editor, *The New Cognitive Neurosciences*, pages 539–552. MIT Press.

Rolls, E. (2003). *The Brain and Emotion*. Oxford University Pres.

Rosenblatt, K. and Payton, D. (1989). A fine-grained alternative to the subsumption architecture for mobile robot control. In *Proceedings of the IEEE/INNS International Joint Conference on Neural Networks. IEEE.*

Ross, J., Morrone, M., Goldberg, M. E., and Burr, D. C. (2001). Changes in visual perception at the time of saccades. *Trends in Neuroscience*, 24:113–121.

Rumelhart, D., Hinton, G., and Williams, R. (1986). Learning internal representations by error propagation. In Rumelhart, D. and McClelland, J., editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume I. MIT Press.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *J. Neurophysiology*, 80:1–27.

Schultz, W., Apicella, P., and Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *Journal of Neuroscience*, (13):900–913.

Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275:1593–1599.

Spier, E. and McFarland, D. (1996). A finer-grained motivational model of behaviour sequencing. In Spier, E. and McFarland, D., editors, *A Finer-Grained Motivational Model of Behaviour Sequencing. From Animals to Animats 4: Proceedings of SAB96.*

Spier, E. and McFarland, D. (1997). Possibly optimal decision-making under self-sufficiency and autonomy. *Journal of Theoretical Biology*, (189):317–331.

St. Amant, R. (1999). User interface affordances in a planning representation. *Human-Computer Interaction*, 14(3):317–354.

Steels, L. (1994). Building agents with autonomous behavior systems. In Steels, L. and Brooks, R., editors, *The 'artificial life' route to 'artificial intelligence'. Building situated embodied agents*. New Haven: Lawrence Erlbaum Associates.

Stephens, D. and Krebs, J. (1986). *Foraging Theory*. Princetown University Press.

Suchman, L. (1987). *Plans and situated action —The problem of human-machine interaction*. Cambridge University Press, Cambridge, UK.

Suonuuti, H. (1997). *Guide to Terminology*. TSK, Nordterm 8, Helsinki.

Suri, R. and Schultz, W. (1998). Learning of sequential movements by neural network model with dopamine–like reinforcement signal. *Exp. Brain Res.*, 121:350–354.

Sutton, R. and Barto, A. (1981). Towards a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, 198(88):135–170.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning*. MIT Press.

Tao, L., Shelley, M., McLaughlin, D., and Shapley, R. (2004). An egalitarian network model for the emergence of simple and complex cells in visual cortex. *PNAS*, (101):366–371.

Thorndike, E. L. (1911). *Animal Intelligence*. The Animal Behaviour Series. New York: Macmillan, p. v.

Tinbergen, N. (1951). *The Study of Instinct*. Oxford University Press.

Tinbergen, N. (1953). *Social Behaviour in Animals*. Methuen's monographs on biological subjects. Methuen.

Toates, F. (1986). *Motivational Systems*. Cambridge University Press.

Toates, F. and Jensen, P. (1990). Ethological and psychological models of motivation - towards a synthesis. In Meyer, J.-A. and Wilson, S. W., editors, *Proceedings of the First International Converence on Simulation of Adaptive Behaviour*, A Bradford Book., pages 194–205. The MIT Press.

Toussaint, M. (2003). Learning a world model and planning with a self-organizing, dynamic neural system. In *Proceedings of NIPS' 2003*.

Tyrrell, T. (1993). *Computational Mechanisms for Action Selection*. PhD thesis, The University of Edinburgh.

Usher, M. and Davelaar, E. J. (2002). Neuromodulation of decision and response selection. *Special Issue on Computational Models of Neuromodulation*, 15(635-645).

Velásquez, J. (1998). Modeling emotion-based decision-making. In *Proceedings of the 1998 AAAI Fall Symposium Emotional and Intelligent: The Tangled Knot of Cognition*, Technical REport FS-98-03, pages 164–169. Orlando, FL: AAAI Press.

Vogt, P. (2001). Adaptive grounding of lexicons on mobile robots. In In de Back, v. d. Z. and Zwanepol, editors, *Robo Sapiens Proceedings of the First Dutch Symposium on Embodied Intelligence*.

von Uexküll, J. (1921). *Umwelt und Innenwelt der Tiere*. Julius Springer, Berlin.

Von Uexkull, J. (1926). *Theoretical Biology*. Kegal Paul, Trench, Trubner & Co. Ltd.

Watkins, C. J. (1989). *Learning from Delayed Rewards*. PhD thesis, Cambridge University.

Westermann, G. and Miranda, E. (2002). Modelling the development of mirror neurons for auditory-motor integration. *Journal of New Music Research*, 31:367–375.

Wilson, S. W. (1991). The animat path to ai. In Meyer, J.-A. and Wilson, S., editors, *From Animals to Animats: Proceedings of the First International Conference on Simulation of Adaptive Behavior*. The MIT Press.

Wright, S. (1931). Evolution in mendelian populations. *Genetics*, (16):97–159.

Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding and selection in evolution. In *Proceedings of the 6th International Congress of Genetics*, number 1, pages 356–366.

Ziemke, T. (1998). Adaptive behavior in autonomous agents. *Presence*, 7(6):564–587.